

# PENTACET data - 23 Million Contextual Code Comments and 250,000 SATD comments

Murali Sridharan  
M3S, University of Oulu  
Oulu, Finland  
murali.sridharan@oulu.fi

Leevi Rantala  
M3S, University of Oulu  
Oulu, Finland  
leevi.rantala@oulu.fi

Mika Mäntylä  
M3S, University of Oulu  
Oulu, Finland  
mika.mantyla@oulu.fi

**Abstract**—Most Self-Admitted Technical Debt (SATD) research utilizes explicit SATD features such as ‘TODO’ and ‘FIXME’ for SATD detection. A closer look reveals several SATD research uses simple SATD (‘Easy to Find’) code comments without contextual data (preceding and succeeding source code context). This work addresses this gap through PENTACET (or 5C dataset) data. PENTACET is a large Curated Contextual Code Comments per Contributor and the most extensive SATD data. We mine 9,096 Open Source Software Java projects totaling over 400 million LOC. The outcome is a dataset with 23 million code comments, preceding and succeeding source code context for each comment, and more than 250,000 SATD comments, including both ‘Easy to Find’ and ‘Hard to Find’ SATD. We believe PENTACET data will further SATD research using Artificial Intelligence techniques.

**Index Terms**—Mining Software Repositories, Technical Debt, Self-Admitted Technical Debt, Team Size

## I. INTRODUCTION

Technical Debt (TD) refers to the cost of rework for having resorted to quicker but sub-optimal solutions in software development [1]. Since 1990, multiple research efforts have investigated TD detection, prioritization, and management techniques. Our focus is on the type of technical debt that is detected through the source code comments left by software developers, called Self-Admitted Technical Debt (SATD) [2]. ‘TODO’ and ‘FIXME’ are tags<sup>1</sup> used by SonarQube [3], one of the widely used static code analysis tools [4] in Continuous Integration/Deployment environment, for technical debt detection through code comments among a host of other code smells. A recent study [5] characterized SATD as ‘Easy to Find’ (can be detected using simple features such as ‘TODO’, ‘FIXME’, ‘XXX’, etc.) and ‘Hard to Find’ (typically refers to multi-word features that do not explicitly indicate SATD, e.g., ‘went horribly wrong’). Identifying and characterizing the ‘Hard to Find’ SATD will benefit software companies with improved SATD detection as a recent survey finds similar SATD comments between Open Source Software (OSS) and industry [6]. Our dataset is the first attempt to characterize the ‘Hard to Find’ SATD as proposed in [5].

Several research efforts [7], [8], [9] for SATD detection utilize Maldonado’s dataset [10]. It utilizes the 64 SATD patterns identified in [2]. This dataset is limited to 10 OSS projects, and the dataset ails with many duplicates, as highlighted

in [11]. Furthermore, past findings from mining Github [12] source code repositories highlight the importance of using probabilistic sampling for dataset construction instead of non-probabilistic sampling in [9], [10] to enable the improved generalization of the findings.

On the human factors front, TD, apart from impacting software productivity adversely also hampers software developers’ morale [13]. It is one of the vital human factors essential in delivering high-quality software [14]. While multiple studies investigate the impact of developer team size (one of the crucial factors of software productivity) on attributes both internal [15], [16] and external to the software [17], [18], the role of the developer team size on SATD and code commenting practices are yet to be explored.

Motivated by these factors and inspired by the study that attempts to understand source code comments at large scale [19], we present PENTACET (or 5C dataset) based on well-defined curation rules. It consists of multiple clusters of bi-directional contextual (i.e., source code preceding and succeeding a code comment) source code comments mined from 9,096 OSS repositories.

We group the clusters of contextual source code comments based on contributor count (developer team size), offering researchers leverage to associate and investigate the source code comments of the repository with the team size (referring to the developer with commit access to an OSS repository). We utilize SoCCMiner [20] to extract the bi-directional code context of a comment. We query the official GitHub REST API<sup>2</sup> to fetch the contributor/s of a repository.

**Contributions:** Our contributions in this dataset are multi-fold:

- 1) first attempt to identify and characterize ‘Hard to Find’ SATD.
- 2) first big SATD dataset with more than 250,000 SATD comments.
- 3) first dataset to capture the bi-directional source code context of SATD comments.
- 4) first dataset to link team size with the contextual source code comments and SATD, to the best of our knowledge.

**Dataset Availability:** Our dataset<sup>3</sup> is made available in the

We thank Academy of Finland (grant ID 328058) for the financial support.

<sup>1</sup>[https://rules.sonarsource.com/java/type/Code Smell?search=tags](https://rules.sonarsource.com/java/type/Code%20Smell?search=tags)

<sup>2</sup><https://api.github.com/repos/OWNER/REPO/contributors>

<sup>3</sup><http://dx.doi.org/10.5281/zenodo.7757462>

open forum [21] to foster a host of deep learning research involving source code comments and SATD. The applications are plentiful from modeling to behavioral and predictive analytics.

## II. METHODOLOGY

The PENTACET data construction pipeline is a multi-step process. Figure 1 depicts the methodology for creating PENTACET data.

**Mining Rules:** The mining rules are tailored to mitigate the risks of mining OSS repositories (repos) identified in [22]. At first, we filter the “Java” OSS repositories with at least 10 stars and, have at least one code push in the past three years since January 01, 2020. It resulted in the initial repository count of 34,545. Then, to exclude non-english repositories, we apply langid<sup>4</sup> tool on the repository description text. “langid” is a python-based language identification stand-alone tool [23]. It identified 24,718 English repositories from 34,545 repositories. Of the 24,718 repositories, we identified 3,101 repositories without any description text. Such repositories without description are excluded resulting in 21,617 repositories for the next filtering criteria.

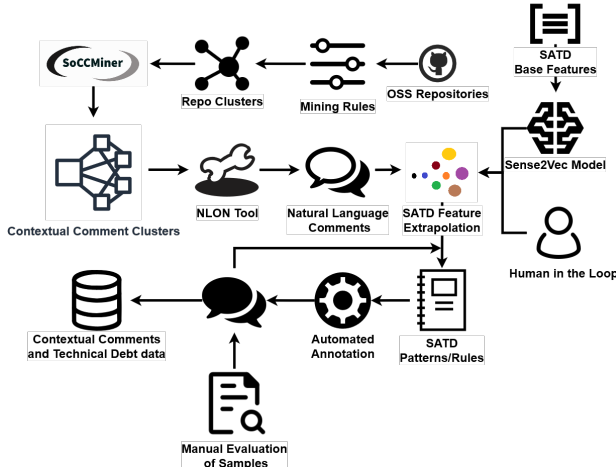


Fig. 1: PENTACET - Data Construction Pipeline

To eliminate the repositories that have educational materials (such as course notes, books, assignments and course materials) and inactive repositories (i.e. repositories that are not maintained and explicitly acknowledged by repository owners), the first author went through the repository (repo) description of the filtered repositories manually and identified 225 (e.g., ‘tutorial’, ‘programming\_assignments’, ‘sample repo’, ‘demonstration project’, ‘textbook’, ‘curriculum’, etc.) patterns for identifying repositories containing explicit educational materials and 31 patterns (e.g., ‘archived’, ‘unmaintained’, ‘outdated’, ‘deprecated’, ‘no longer supported’, ‘obsolete’, ‘discontinued’, etc.) to identify explicit non-maintenance repositories respectively. Among the 21,617 repositories, 6,813

<sup>4</sup><https://github.com/saffsd/langid.py>

are identified as educational related and 244 repositories are identified as non-maintenance repositories, respectively and are excluded. It is followed by filtering repositories (a total of 202 repositories) with repository names that indicate educational or non-maintenance related repositories. For example, ‘AndroidDemoProjects’, ‘Azure-Samples’, ‘TFLiteFaceExample’, etc., Now, there are 14,358 repositories for further processing. Lastly, the repositories (3,718) that are archived and without any license are removed, with 10,640 repositories progressing to next curation. Finally, to capture the active repositories, we order the repositories based on the number of repository commits [22] since January 01, 2020 in descending order and consider only those repositories that have at least 3 commits since January 01, 2020, assuming at least one commit a year. This leaves us 9,111 repositories. After removing 15 invalid repositories with ‘.javascript’ files instead of ‘.java’, a total of 9,096 repositories move to the next stage in the dataset construction pipeline.

**Repository Clusters:** Each OSS repository in GitHub has ‘Contributors’ attribute comprising both core and co-developers [24]. We cluster the curated repositories based on ‘Contributor Count’ that ranges between 1 and 30 for the 9,096 repositories. We split the repositories into 4 clusters. The number of clusters are purely for organizational convenience, while any new clusters can be formed or rearranged by querying PENTACET database using ‘CONTRIBUTOR\_COUNT’ from ‘PROJECT\_MAIN’ table as discussed in Section III. Cluster 1 (C1) contains repositories of single contributor, most likely the author of the repository. We group repositories with contributor count 2 and 3 in Cluster 2 (C2), contributor count 4 till 10 in Cluster 3 (C3) and contributors greater than 10 in Cluster 4 (C4).

**SoCCMiner:** We employ SoCCMiner [20] to extract the contextual information of source code comment along with the comprehensive project attributes (including project, file, package, class, interface, enum, and, static block attributes). The essence of SoCCMiner is its ability to extract bi-directional contextual source code of a comment. For example: consider the comment “TODO elastic?” in Table I. The unique capability of SoCCMiner apart from fetching parent trace for code comments is that it fetches both preceding and succeeding code context. It senses one of the special case in which the comment is in the first line of a function, and therefore to fetch the preceding code context, it fetches the entire function. Such bi-directional code context is crucial for code comprehension, comment generation of non-header comments.

TABLE I: Code Context Extraction Comparison

Context in prior work [25]	Code Context in SoCCMiner [20]
// TODO elastic? rows.add(ix, val);	<pre>public void NoEvent(int ix, V val){ // TODO elastic? rows.add(ix, val); }</pre>

**NLoN:** Source code comments often contain commented out source code, or other non-natural language elements. To

remove code comments that do not contain natural language text, we utilize NLoN [26] tool that accept code comments and perform a binary classification if the comment is natural language ('NL') or not ('Not'). The results are updated in PENTACET database for every comment under 'NLON\_STATUS' column of the COMMENT\_ATTR table.

**SATD Feature Extrapolation:** We generate new SATD features based on the existing 64 established SATD features in [2] with Sense2Vec [27]. Sense2Vec captures contextually similar words with its deep word embedding. For example, for feature 'problematic', Sense2Vec returns multiple semantically similar words such as 'troublesome', 'undesirable', 'contentious', 'separate issue', etc.,. We use these extrapolated features (words) to evaluate our code comments for SATD.

TABLE II: PENTACET (5C Dataset) Stats

	CLUSTER 1 (#1 Contributor)	CLUSTER 2 (#2-3 Contributors)	CLUSTER 3 (#4-10 Contributors)	CLUSTER 4 (> #10 Contributors)	TOTAL
# Projects	1,576	2,202	3,004	2,314	9,096
# Comments	1,055,517	2,498,870	4,660,481	15,079,741	23,294,609
# Invalid Comments	31,118	37,778	84,876	152,193	305,965
# NL Comments	733,409	1,737,444	3,281,562	10,336,159	16,088,574
# SATD Comments	23,039	54,052	73,204	329,506	479,801
% SATD Comments	3.14	3.11	2.23	3.18	2.98

Sense2Vec extrapolated 'Easy to Find' features such as ('sabotaging', 'damaging', 'counter-productive', etc.,) are passed to the first round of automated SATD annotation. The first author manually evaluates the results by reading the comments. The well-established 'Easy to Find' [5] features such as 'TODO', 'FIXME', and 'XXX' were excluded in the manual validation. During the validation process, the first author creates the heuristics for 'Hard to Find' [5] SATD. For example, the 'Hard to Find' SATD feature 'not sure if this is' returns results as the following:

```
//not sure if this is really what should be done
//not sure if this is the right thing to do.
//not sure if this is needed?
//not sure if this is correct, double check!!
//not sure if this is required, but looks feasible
```

This iterative process involved multiple iterations of annotation, manual evaluation, creating heuristics (pattern matching), and creating features for the 'Hard to Find' SATD.

**SATD Annotation:** Features that attract too many false positives are removed from the SATD feature list. For example, 'cause problem' will capture the comment '// Empty bboxes can cause problems' which sound right at first. Later, after manual scanning, several header comments had the following comment content '//This won't cause problems'. Such features are removed from the SATD feature list to avoid overfitting features. We choose C1 cluster for heuristics' creation since we wanted to start with the smallest (in terms of #comments) considering the time and manual effort involved in the annotation.

It took two months for the first author to complete the SATD feature annotation. We have included both 'Easy to Find' and 'Hard to Find' SATD features in [21]. Table II contains the SATD split for each cluster. To the best of our knowledge, this is the first attempt to explore and annotate the 'Hard to Find' SATD. We have characterized 655 'Hard to Find' SATD features in PENTACET [21]. Although this process is semi-automated, involving a human annotator, we believe this data will be the foundation for much more sophisticated, completely automated techniques for 'Hard to Find' SATD detection.

For evaluating the quality of the dataset, we obtained a stratified random (balanced sample strength to accommodate both 'SATD' and 'Non-SATD') and a statistically significant sample size 385<sup>5</sup> accounting for 95% confidence level with 5% margin for error. The first author manually validated 1,540 samples. To remove any bias in the annotation process, we took a random sample of 97 from each cluster, and the second author annotated a total sample size of 388. The second author is a doctoral researcher whose primary area of research is Technical Debt. The Inter-Rater Agreement (IRA) is evaluated through Cohen's Kappa Co-efficient. The IRA for our dataset achieved  $k=0.75$ , which is substantial [28], [29].

### III. DATA MODEL

We use PostgreSQL<sup>6</sup> 14.0 database to store the PENTACET schema. All the project meta information are stored in the 'PROJECT\_MAIN' table as observed in Figure 2.

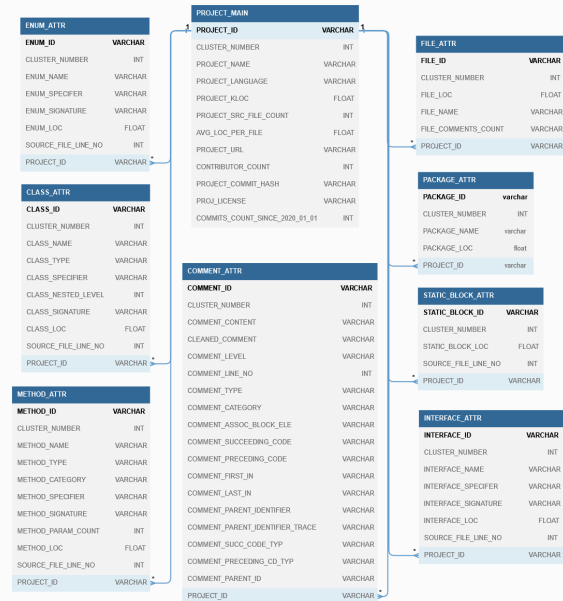


Fig. 2: PENTACET Schema

The database dump file for the PENTACET contains more than 20 million bi-directional contextual code comments along with comprehensive project attributes as in [20].

<sup>5</sup><https://www.surveymonkey.com/mp/sample-size-calculator>

<sup>6</sup><https://www.postgresql.org/>

The attributes of each granularity in a project are stored in their respective attribute table. For example, file attributes are stored in ‘FILE\_ATTR’ table, method attributes are stored in ‘METHOD\_ATTR’ table, etc., The main table containing a bulk of data is ‘COMMENT\_ATTR’ table. The ‘COMMENT\_ATTR’ table contains ‘NLON\_STATUS’ column that contains either ‘NL’ (Natural Language) or ‘Not’ (Not Natural Language) or ‘INVALID’ (empty comments after removing any decorative comment symbols such as ‘///’, ‘—’, ‘\*\*\*’, ‘###’, and trimming trailing and leading whitespaces). The ‘SATD\_AFFLICTION’ column contains the SATD status of the comment and ‘SATD\_FEATURE’ column contains the SATD features that caused the comment to be SATD in the annotation process.

For illustration purposes, the query i) fetches the SATD afflicted natural language comments from those projects whose Lines of Code (LOC) is greater than 10,000 and can be from clusters 1 to 4 based on the input.

```
i) SELECT COMMENT_ID, COMMENT_CONTENT
FROM COMMENT_ATTR
WHERE CLUSTER_NUMBER=[1-4] AND
PROJECT_ID IN (SELECT PROJECT_ID
FROM PROJECT_MAIN WHERE PROJECT_LOC > 10000)
AND NLON_STATUS = 'NL'
AND SATD_AFFLICTION= 'SATD'
```

Similarly, comments can be fetched based on the other granular (class, method, interface, etc.,) attributes.

#### IV. APPLICATIONS

**Modeling:** Modeling code comments with more than 16 M natural language comments with code contextual information and almost 0.5 M SATD comments will aid deep learning for source code comments and technical debt research. The extended attributes of all project granularities associated with the contextual code information enable large-scale source code characterization for SATD.

**Predictive Analytics:** The characterization of SATD in conjunction with project size (in LOC), file size (in LOC), other granular code constructs’ (such as class, method, interface, etc.,) size (in LOC), and other attributes generated by SoCCMiner [20] can help investigate its relation with SATD enabling correlational analysis and forecasting applications.

**Behavioural Analytics:** The contributor mapping to the contextual source code metrics can help investigate developers’ commenting practices across different project sizes, e.g., are there more or fewer comments in large projects? It can further help study sentiment analysis and assist with correlation analysis of developers’ commenting practices with SATD.

#### V. ETHICAL CONSIDERATIONS

We carefully considered ensuring we adhere to the ethical mining guidelines as in [30]. In particular, repositories with no license attached to them were ignored and not included in the dataset. We carefully monitored to anonymize any personal information (author\_id of the repository) while processing PENTACET data.

#### VI. THREATS TO VALIDITY

**Internal Validity:** The primary limitation of this dataset is that the SATD patterns identified during the iterative process are taken from cluster 1 (C1) alone. However, this is a step towards capturing the fine-grained features of SATD that can enable deep learning techniques to extract such data from unseen data. The comments might contain noise, even in comments categorized as natural language. Data is prone to be noisy; careful curation and initial efforts toward automating noise removal are paramount.

**External Validity:** Although this is huge, open source data, it is difficult for this dataset to characterize SATD features exhaustively or claim generalization of SATD features across all the projects. However, this work is the largest SATD dataset to our knowledge.

#### VII. RELATED WORK

The Code and Comments dataset [25] and CoDesc [31] dataset captures code comment pairs. Comparatively our data is larger and in addition, SoCCminer supplements bidirectional code context information at all granular level (class, method, interface, etc.,). Such context information will be very crucial for identifying the locale or code context of non-header comments. The Technical Debt dataset [32] primarily concentrates on TD generated by SonarQube [3], while PENTACET characterizes ‘Hard to Find’ SATD present in source code comments. 20-MAD [33] dataset focuses on comments from issues and commits while our dataset focuses on source code comments along with source code contextual information.

#### VIII. CONCLUSION

We apply extensive curation for mining OSS repositories and present the largest bi-directional code contextual data of 23 M source code comments at all source code granularities for 9096 Java projects. Further, we fetch and associate each repository’s contributor count, enabling correlational and characterization analysis relating to source code comments and developer team size. Thirdly, we have compiled the largest SATD dataset through a semi-automated annotation process after removing non-natural language content (source code) and noise (empty comments or dummy comments with just symbols) using the NLoN tool. In the process, we extended the SATD features from [2] with Sense2Vec [27] and added manually verified features that detect ‘Hard to Find’ SATD as suggested in [5]. We achieved a substantial agreement in the inter-rater agreement during manual validation.

For future work, we plan to fetch the source code, static analysis, git process metrics for all the repositories in PENTACET and associate SATD comments with the generated metrics using the unique combination of comment parent identifier and comment trace (SoCCMiner [20] attributes). It will enable software researchers to investigate the correlational study and improve SATD forecasting.

## REFERENCES

- [1] W. Cunningham, "The wycash portfolio management system," *ACM SIGPLAN OOPS Messenger*, vol. 4, no. 2, pp. 29–30, 1992.
- [2] A. Potdar and E. Shihab, "An exploratory study on self-admitted technical debt," in *2014 IEEE International Conference on Software Maintenance and Evolution*. IEEE, 2014, pp. 91–100.
- [3] SonarSource, "Sonarqube," 2006. [Online]. Available: <https://www.sonarsource.com/products/sonarqube/>
- [4] C. Vassallo, F. Palomba, A. Bacchelli, and H. C. Gall, "Continuous code quality: Are we (really) doing that?" in *Proceedings of the 33rd ACM/IEEE International Conference on Automated Software Engineering*, 2018, pp. 790–795.
- [5] Z. Yu, F. M. Fahid, H. Tu, and T. Menzies, "Identifying self-admitted technical debts with jitterbug: A two-step approach," *IEEE Transactions on Software Engineering*, 2020.
- [6] F. Zampetti, G. Fucci, A. Serebrenik, and M. Di Penta, "Self-admitted technical debt practices: a comparison between industry and open-source," *Empirical Software Engineering*, vol. 26, no. 6, pp. 1–32, 2021.
- [7] Q. Huang, E. Shihab, X. Xia, D. Lo, and S. Li, "Identifying self-admitted technical debt in open source projects using text mining," *Empirical Software Engineering*, vol. 23, pp. 418–451, 2018.
- [8] X. Ren, Z. Xing, X. Xia, D. Lo, X. Wang, and J. Grundy, "Neural network-based detection of self-admitted technical debt: From performance to explainability," *ACM transactions on software engineering and methodology (TOSEM)*, vol. 28, no. 3, pp. 1–45, 2019.
- [9] Z. Guo, S. Liu, J. Liu, Y. Li, L. Chen, H. Lu, and Y. Zhou, "How far have we progressed in identifying self-admitted technical debts? a comprehensive empirical study," *ACM Transactions on Software Engineering and Methodology (TOSEM)*, vol. 30, no. 4, pp. 1–56, 2021.
- [10] E. d. S. Maldonado and E. Shihab, "Detecting and quantifying different types of self-admitted technical debt," in *2015 IEEE 7th international workshop on managing technical debt (MTD)*. IEEE, 2015, pp. 9–15.
- [11] M. Sridharan, M. Mantyla, L. Rantala, and M. Claes, "Data balancing improves self-admitted technical debt detection," in *2021 IEEE/ACM 18th International Conference on Mining Software Repositories (MSR)*. IEEE, 2021, pp. 358–368.
- [12] V. Cosentino, J. Luis, and J. Cabot, "Findings from github: methods, datasets and limitations," in *Proceedings of the 13th International Conference on Mining Software Repositories*, 2016, pp. 137–141.
- [13] T. Besker, H. Ghanbari, A. Martini, and J. Bosch, "The influence of technical debt on software developer morale," *Journal of Systems and Software*, vol. 167, p. 110586, 2020.
- [14] E. Guveyi, M. S. Aktas, and O. Kalipsiz, "Human factor on software quality: a systematic literature review," in *International Conference on Computational Science and Its Applications*. Springer, 2020, pp. 918–930.
- [15] P. S. Kochhar, T. F. Bissyandé, D. Lo, and L. Jiang, "An empirical study of adoption of software testing in open source projects," in *2013 13th International Conference on Quality Software*. IEEE, 2013, pp. 103–112.
- [16] A. Youssef and A. Capiluppi, "The impact of developer team sizes on the structural attributes of software," in *Proceedings of the 14th International Workshop on Principles of Software Evolution*, 2015, pp. 38–45.
- [17] S. D. Vishnubhotla, E. Mendes, and L. Lundberg, "Understanding the perceived relevance of capability measures: A survey of agile software development practitioners," *Journal of Systems and Software*, vol. 180, p. 111013, 2021.
- [18] S. Tyagi, R. Sibal, and B. Suri, "Empirically developed framework for building trust in distributed agile teams," *Information and Software Technology*, vol. 145, p. 106828, 2022.
- [19] H. He, "Understanding source code comments at large-scale," in *Proceedings of the 2019 27th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, 2019, pp. 1217–1219.
- [20] M. Sridharan, M. Mäntylä, M. Claes, and L. Rantala, "Socminer: a source code-comments and comment-context miner," in *Proceedings of the 19th International Conference on Mining Software Repositories*, 2022, pp. 242–246.
- [21] M. Sridharan, L. Rantala, M. Mäntylä, and M. Claes, "Pentacet data - 23 million contextual code comments and 500,000 satd comments," M3S, University of Oulu, Oulu, Finland, 2023. [Online]. Available: <https://github.com/M3SOulu/pentacet>
- [22] E. Kalliamvakou, G. Gousios, K. Blincoe, L. Singer, D. M. German, and D. Damian, "An in-depth study of the promises and perils of mining github," *Empirical Software Engineering*, vol. 21, no. 5, pp. 2035–2071, 2016.
- [23] M. Lui and T. Baldwin, "Cross-domain feature selection for language identification," in *Proceedings of 5th International Joint Conference on Natural Language Processing*. Chiang Mai, Thailand: Asian Federation of Natural Language Processing, Nov. 2011, pp. 553–561. [Online]. Available: <https://aclanthology.org/I11-1062>
- [24] N. Matragkas, J. R. Williams, D. S. Kolovos, and R. F. Paige, "Analysing the 'biodiversity' of open source ecosystems: the github case," in *Proceedings of the 11th working conference on mining software repositories*, 2014, pp. 356–359.
- [25] B. Gelman, B. Obayomi, J. Moore, and D. Slater, "Source code analysis dataset," *Data in brief*, vol. 27, p. 104712, 2019.
- [26] M. V. Mäntylä, F. Calefato, and M. Claes, "Natural language or not (nlon) a package for software engineering text analysis pipeline," in *Proceedings of the 15th International Conference on Mining Software Repositories*, 2018, pp. 387–391.
- [27] A. Trask, P. Michalak, and J. Liu, "sense2vec-a fast and accurate method for word sense disambiguation in neural word embeddings," *arXiv preprint arXiv:1511.06388*, 2015.
- [28] J. L. Fleiss, B. Levin, M. C. Paik *et al.*, "The measurement of interrater agreement," *Statistical methods for rates and proportions*, vol. 2, no. 212–236, pp. 22–23, 1981.
- [29] K. E. Emam, "Benchmarking kappa: Interrater agreement in software process assessments," *Empirical Software Engineering*, vol. 4, no. 2, pp. 113–133, 1999.
- [30] N. E. Gold and J. Krinke, "Ethics in the mining of software repositories," *Empirical Software Engineering*, vol. 27, no. 1, pp. 1–49, 2022.
- [31] M. Hasan, T. Muttaqueen, A. A. Ishtiaq, K. S. Mehrab, M. Haque, M. Anjum, T. Hasan, W. U. Ahmad, A. Iqbal, and R. Shahriyar, "Codesc: A large code-description parallel dataset," *arXiv preprint arXiv:2105.14220*, 2021.
- [32] V. Lenarduzzi, N. Saarimäki, and D. Taibi, "The technical debt dataset," in *Proceedings of the Fifteenth International Conference on Predictive Models and Data Analytics in Software Engineering*, 2019, pp. 2–11.
- [33] M. Claes and M. V. Mäntylä, "20-mad: 20 years of issues and commits of mozilla and apache development," in *Proceedings of the 17th International Conference on Mining Software Repositories*, 2020, pp. 503–507.