# An Empirical Study of License Conflict in Free and Open Source Software

# 大规模开源软件许可证冲突风险分析实证研究

崔星 吴敬征 武延军 王旭 罗天悦 屈晟 凌祥 杨牧天

The 45th International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP 2023)

联系人：崔星，13051316652，cuixing@iscas.ac.cn

## Background

**Free and Open Source Software (FOSS) has become the fundamental infrastructure of mainstream software projects :**

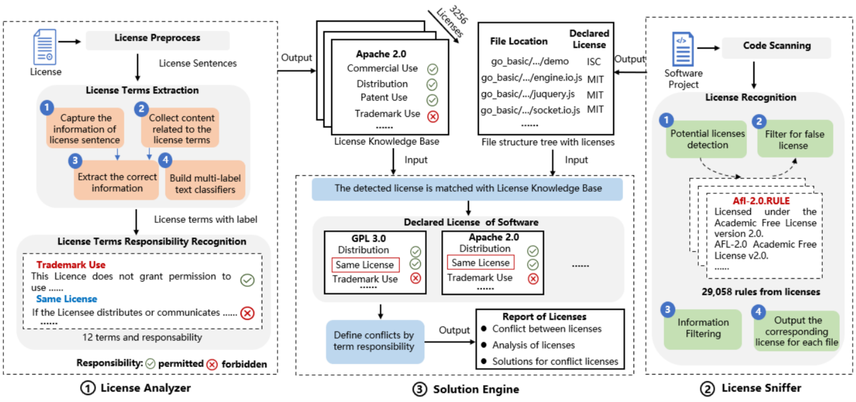
- FOSS has been widely adopted by academia and industry and extensively deployed by commercial products in their production. The redistribution and commercialization of new software will be negatively impacted if the contradictory terms or violations are found within the FOSS.

**Large number of license compatibility issues in the open source software supply chain:**

- Sometimes developers will combine multiple components under different open source licenses with direct or transitive dependencies, raising the legal concern of whether their licenses are compatible.

## Methodology

We design DIKE for conflict detection and resolution across a broad range of licenses, without being restricted by the project's programming language.




## Evaluation

Table 1: Performance comparison of different models in license terms extraction

| Model | macro Precision | macro Recall | macro F1-Score |
|---|---|---|---|
| ALBERT | 0.839 (±0.021) | 0.765 (±0.035) | 0.782 (±0.027) |
| ALBERT+Focal Loss | 0.883 (±0.026) | 0.755 (±0.014) | 0.795 (±0.014) |
| LSAN | 0.806 (±0.019) | 0.756 (±0.023) | 0.767 (±0.016) |
| LSAN+Focal Loss | 0.886 (±0.018) | 0.766 (±0.022) | 0.816 (±0.014) |

Table 2: Performance of models in license terms responsibility recognition

| Model | Accuracy | Precision | Recall | F1-Score |
|---|---|---|---|---|
| ALBERT | 0.943 (±0.008) | 0.937 (±0.014) | 0.960 (±0.011) | 0.948 (±0.007) |

Table 3: Performance comparison with manual analysis

| ID | Manual | DIKE | Accuracy | Precision | Recall | F1-Score |
|---|---|---|---|---|---|---|
| 1 | 258 | 243 | 194 (75.2%) | 0.798 | 0.752 | 0.774 |
| 2 | 263 | 257 | 196 (74.5%) | 0.763 | 0.745 | 0.754 |
| 3 | 264 | 276 | 213 (80.7%) | 0.772 | 0.807 | 0.789 |
| AVG | 262 | 259 | 201 (76.7%) | 0.778 | 0.768 | 0.772 |

- Our License Analyzer model achieves a macro F1-score of 0.816, representing a 3.4%, 2.1%, and 5.5% improvement compared to ALBERT, ALBERT+Focal Loss, and LSAN, respectively.

## Conclusion

- DIKE analyzes 3,256 open source licenses to build a knowledge base, aiding in conflict detection and resolution across a broad range of licenses. DIKE offers two solutions to resolve these license conflicts: (1) Replace the open source license; (2) Replace the source code.
- We analyze 16,341 popular open source projects on GitHub, using 1,787 different licenses, and found that 27.2% (4,448) had license conflicts. This highlights the significant issue of improper license usage in a large number of projects.

## Result and Analysis

Table 4: Statistics of the license conflict

| License Conflict Pair | Project Number | Propertion |
|---|---|---|
| MIT with GPL-3.0-or-later | 393 | 2.40% |
| MIT with GPL-2.0 | 372 | 2.27% |
| Apache-2.0 with GPL-3.0-or-later | 232 | 1.41% |
| Apache-2.0 with GPL-2.0 | 227 | 1.38% |
| MIT with GPL-3.0 | 199 | 1.20% |
| MIT with GPL-2.0-or-later | 169 | 1.03% |
| GPL-3.0 with GPL-2.0 | 146 | 0.89% |
| BSD-3-Clause with GPL-3.0-or-later | 138 | 0.84% |
| Apache-2.0 with MPL-2.0 | 125 | 0.76% |
| GPL-2.0 with GPL-3.0-or-later | 125 | 0.76% |

Table 5: Statistics on the use of open source license terms

| License terms | Conflict Project | Propertion |
|---|---|---|
| Same license | 4,148 | 25.38% |
| Patent use | 2,643 | 16.17% |
| Disclose source | 2,087 | 12.77% |
| Commercial use | 407 | 2.49% |
| License and copyright notice | 208 | 1.27% |
| Private use | 150 | 0.92% |
| Trademark use | 82 | 0.50% |
| Distribution | 27 | 0.17% |
| Modification | 4 | 0.02% |
| State changes | / | / |
| Liability | / | / |
| Warranty | / | / |

- We collect 16,341 FOSS projects with more than 1,000 stars on GitHub. Approximately half of the disputes result from combinations of the GPL family and open source licenses.
- The terms used in the licenses of the conflicting projects are associated with 'Same license, Patent use, and Disclose source'.

## Application

- DIKE has been used by Beijing ZhongKeWeiLan Technology Co.,Ltd., and is serving several large-scale Internet and government enterprises, which shows that DIKE has certain industrial value.

This paper is supported by Open Source Map