Application of Computational Linguistics Techniques for Improving Software Quality*

Amin Boudeffa[1], Antonin Abherve[1], Alessandra Bagnato[1], Cedric Thomas[2], Martin Hamant[2], and Assad Montasser[2]

[1] Softeam, Paris, France
Email: *firstname.lastname*@softeam.fr
[2] OW2, Paris, France
Email: *firstname.lastname*@ow2.org

Abstract. Progress in Artificial Intelligence, Big Data and Computational Linguistics domains offered new way to perform n-depth analysis and evidence-based quality assessments of open source software components.In this paper we will see how this can be integrated into industrial development to improve the quality of developed software.

Keywords: Computational Linguistics · Big Data · Sentiment analysis

1 Project Data

Developing new software systems by reusing existing open source software (OSS) components raises challenges related to the level of quality of different OSS as well as to the level of support that different OSS communities provide to users of the software they produce[2]. The CROSSMINER project aim to adress this issue.

- **Acronym:** CROSSMINER
- **Title:** Developer-Centric Knowledge Mining from Large Open-Source Software Repositories
- **Start date:** January 1, 2017
- **Duration:** 36 months
- **Partners:** The Open Group, University of L´Aquila, University of York, Softeam, OW2 Consortium, Edge Hill University, Unparallel Innovation, Eclipse Foundation Europe, Centrum Wiskunde & Informatica, Castalia Solutions, Bitergia, Athens University of Economics & Business.
- **Web site:** https://www.CROSSMINER.org/

* Supported by the European Unions Horizon 2020 Research and Innovation Programme.

2 CROSSMINER Analysis Platform

2.1 CROSSMINER Project

CROSSMINER is an EU-funded research project which aims to deliver an integrated open-source platform that will support the development of complex software systems by enabling the monitoring, in-depth analysis and evidence-based selection of open source components, and facilitating knowledge extraction from large open-source software repositories. The project leverages multi-disciplinary sub-fields of computer science including Artificial Intelligence, Big Data and Computational Linguistics. The project aimed six main scientific and technology objectives among which the following four were used in the context of this experimentation :

- Development of source code analysis tools to extract and store actionable knowledge from the source code of a collection of open-source projects
- Development of natural language analysis tools to extract quality metrics related to the communication channels, and bug tracking systems of OSS projects by using Natural Language Processing and text mining techniques
- Development of workflow-based knowledge extractors that simplify the development of bespoke analysis and knowledge extraction tools shielding engineers from technological issues to concentrate on core analysis tasks
- Development of advanced integrated development environments that will allow developers to adopt the CROSSMINER knowledge base and analysis tools directly from the development environment will help developers to improve their productivity.

2.2 Natural Language Processing Metrics

Natural language contains vital and potentially hidden information that can be exploited to assist developers in making vital decisions surrounding open source software development[3]. The natural language components developed within the CROSSMINER project used to analyse various source of information of given Open Source software projects. The NLP metrics compute heuristics that summarise the quality of support offered to users over time, and contribute to the CROSSMINER knowledge base by enriching documents with extra information.[3].

These metrics process the output classification values or a conversion of texts provided by the main basic metrics associated with various natural language tools integrated into CROSSMINER. We distinguish Sentimental Metric that reveals which sentiments are expressed in a bug tracking system for a project and Emotional Metric that Summarises the emotions expressed in the bug tracking systems of a given project.

The state of the art industrial software development process is based on monitoring the product quality through the use of a low level code-based metrics which are related particularly to the software development implementation

phase. In CROSSMINER, the use of NLP tools is of high relevance, as the analysis of the text written by developers and users provides information that would be expensive and laborious to process manually. The extraction of these information about the quality of support offered by the community of an open source software project to be made available the sentiment analysis, classification of emotions, detection of request and replies among messages posted in a communication channel, bug tracker or forum, categorization of messages according to their content type and the classification of threads of messages according to the severity of the issue that they express.

3 Use Case description

3.1 Softeam Use Case

The first company which perform this experimental integration, Softeam, is a French Company of 1000 employees, which operates in many different domains such as Finance, Banking, Insurance and Service industries. The company led this experiment in the context of the development team of a commercial long live software : Modelio, a modelings tool for developers and architects to support software and system engineering.

Each Modelio release follows a specific development process based on the Agile methodology in order to align Modelio features to market demands and guarantee the product quality. Each developments projects start by an initials specification phases in which the perimeter of the release is defined. At the end of each sprint, the quality issues of delivered components are assessed by the quality team by performing validations activity. Feedback are used to modify and adapt the next sprint plan. The quality assurance process can lead to an update of the project plan and require an adaptation of the architecture of the solution, the specification of the features being implemented or the perimeter of the release.

To develop its solutions, Softeam relies more and more frequently on open source libraries. Due to the critically of open source libraries and framework used as core components of his products, the selection and assessment of the quality of these libraries follow the same level of quality evaluation as Softeam internally developed code source.The selection process and administration of this components are a long and costly process and we expect that CROSSMINER will help us to conduct it.

3.2 OW2 Use case

OW2 is a global open source software non-profit association, its mission is to foster the development of a portfolio of open source software for information systems and the growth of a business ecosystem around it. OW2 promotes a code base of some 100 open source projects; its global community membership involves some 40 members, including commercial, public and academic organisations, and over 2500 individual members, half of them from Europe.

As the organisation becomes a reference community platform in the open source marketplace, it increasingly stresses the quality and market readiness of its software. OW2 endeavours to integrate solutions helping projects to produce assessment reports on the quality of the code and on the maturity of their governance.

The OW2 use case has two goals. The first one is to provide project leaders and users with cutting edge tools for analysing and measuring accurately their software information sphere. The second goal is to develop a Market Readiness Index that will helps conventional managers select OW2 projects according to criteria spanning from technology quality to business sustainability.

As a result, OW2 will differentiate from comparable organisations, such as the Linux Foundation and the Eclipse Foundation (also a partner in the CROSSMINER project), which are also working on systems to collect data about their projects.

4 Experimentation

4.1 Increasing quality of Softeam product by including sentiment analysis technics in development process

To increase his capacity to evaluate the quality of open source components embeded in his products, Softeam has integrated the provided solution,including the sentiment analysis and classification of emotions techniques, with his standardized development process.

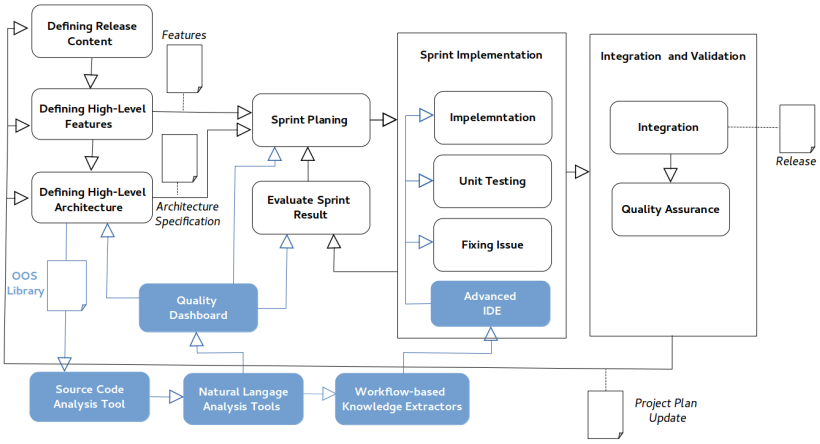

Fig. 1. Softeam standardized development approach with CROSSMINER solution.

In project initiation phase, Softeam evaluates how the source code analysis tool and the natural language analysis tool could be used to assist architects

to choose the open sources framework which will be included in project architecture in order to add new services and functionality in Modelio.The result of sentimental analysis of textual data sources related to the component is used to evaluate how the open source community is reacting towards the specific library. In sprint implementation phases, by the intermediary of the IDE, the Computational Linguistics Techniques to identify the more relevant information that must be delivered to the developers.

The first evaluation of the impact of deploying the solution Softeam showed a significant improvement when working with new open source libraries :

- Reduction of 40% of average time needed to evaluate existing open source components used in a Modelio project architecture.
- Reduction of 25% of average time needed to choose open source components to be included in a project architecture.
- Reduction of 10% of average time for development which involved the use of new libraries unknown to our developers.

4.2 OW2 Experimentation with Sentiment analysis metrics

The experimentation with sentiments fits with OW2s business need to to integrate into its process innovative ways to assess the market readiness of its projects, and to provide project leaders with tools and methods to help them to progress on the path toward greater maturity. The OW2 experimentation concentrates on contributor metrics to provide project leaders with the ability to better monitor and understand the behavior of their contributors.

The first objective is achieved by developing sentiment and emotion analysis based on the application of Natural Language Processing techniques on informal sources such as documentation and code and bug comments. There are three main challenges here. One is to identify metrics that can be collected throughout the whole code base so the method is applicable to all the projects. the second one is to develop data collectors, or readers, that can address the variety of sources. The third challenge is to define how to compute a snapshot indicator from time series covering periods from one quarter to a whole year.

Project	Emotions (count)					
	Surprise	Joy	Love	Sadness	Anger	Fear
XWIKI	326	357	361	364	364	359
Sat4j	0	286	286	184	107	0
asm	5	38	38	37	38	0

Table 1. Sample of the emotions apparition which appear on three projects.

The second objective is addressed by setting up visual user interfaces reflecting the metrics that will get computed based on the tools delivered by CROSS-

MINER. Such visual interfaces will let the user browse both high level and fine grained information, depending on the type of question. One key challenge here is to produce visual representations that are easily understandable by any reader and operationally meaningful for project leaders.

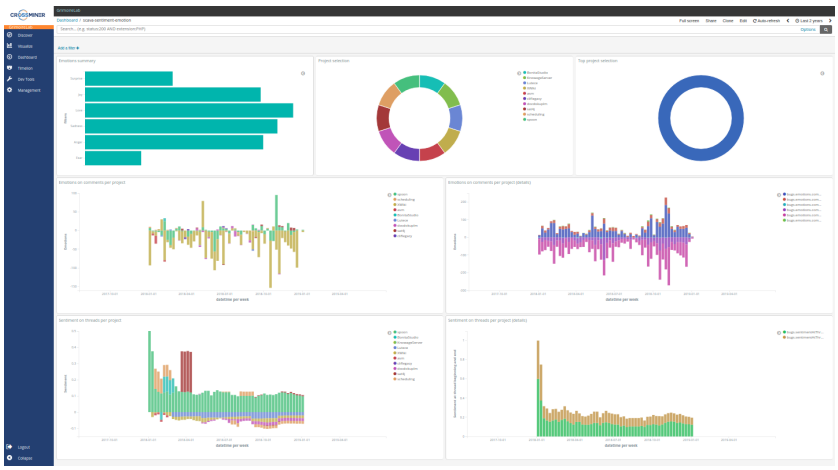

Fig. 2. Dashboard of Sentimental analysis natural language metrics applied so far by OW2 to assess projects

5 Acknowledgments

The research described has been carried out as part of the CROSSMINER Project, which has received funding from the European Unions Horizon 2020 Research and Innovation Programme under grant agreement No. 732223.

References

1. Amin Boudeffa, Alessandra Bagnato, Antonin Abherve, Davide Di Ruscio, Marcio Mateus and Bruno Almeida: Integrating and deploying heterogeneous components by means of a microservice architecture in the CROSSMINER project.STAF-CE **1**(5), 61–66 (2019)
2. Alessandra Bagnato, Konstantinos Barmpis, Nik Bessis, Luis Adrin Cabrera-Diego, Juri Di Rocco, Davide Di Ruscio, Tams Gergely, Scott Hansen, Dimitris S. Kolovos, Philippe Krief, Ioannis Korkontzelos, Stéphane Laurière, Jose Manrique Lopez de la Fuente, Pedro Mal, Richard F. Paige, Diomidis Spinellis, Cedric Thomas, Jurgen J.Vinju: Developer-Cen-tric Knowledge Mining from Large Open-Source Software Repositories (CROSSMINER). STAF Workshops 2017: 375-384 Marburg (2017).
3. Edge Hill University: D3.4 Natural Language Components, 27 December 2017 Final.