SAINSTECH
INNOVATION JOURNAL

SIJ, Volume 3, Nomor 2, November 2020, Hal. 87-95

Feature Extraction Technique For Text Mining Requirement For Reuse in Software Product Lines: A Systematic Literature Review

Fahmi Syuhada, Ramadhana Agung Pratama[1)]

[1)] Computer Science Faculty of Science and Technology, University of Qamarul Huda Badaruddin[2)]

**Corresponding Author: fahmi.uniqhba@gmail.com, Tel: +6281770095288*

Diterima pada *2 Pebruari 2018*, **Direvisi pertama pada** *15 Maret 2018*, **Direvisi kedua pada** *28 Maret 2018*, **Disetujui pada** *22 April 2018*, **Diterbitkan daring pada** *20 Mei 2018*

Abstract: *Requirements for a system implementation can be done by feature extraction process to support the implementation process of system products. Extraction of a data or variable feature from the requirements has the effect of benefiting from Software-Product-Line. Text mining is a process of analyzing a text to generate a product. The Requirements for feature data that is processed in the text mining process can be said to be part of the Product Line Software. There are various kinds of techniques used to produce data feature requirements that enter the text mining process. This paper provides a Systematic Literature Review (SLR) of how feature extraction techniques used in the text mining data feature extraction process to support the Software Product Line process. The author uses literature search techniques to support the process of making this SLR. Besides discussing the feature extraction technique, this SLR also discusses the tools used for the Text-mining Feature Extraction process. This SLR is expected to contribute to provide a review regarding the feature of text mining data extraction techniques to support the Software Product Line process.*

Keywords: *Requirement, Extraction Feature, Software Product Line, Systematic literature review*

1. INTRODUCTION

Software-product-line is a set of software-intensive systems that share a common, managed set of features satisfying the specific needs of a particular market segment or mission and that are developed from a common set of core assets in a prescribed way[11]. New products can be built by reusing software components to obtain benefits like a shorter time-to-market and higher quality, provided that these have already been developed, tested and used. Not only code can be reused, as other assets like test cases, requirements, architectural design, etc., can also be reused [10].

Text mining can be a solution to finding a problem solving from an electronic text information data. Due to the increasing number of electronic information available (digital libraries, electronic mail, and blogs), text mining getting more importance. Text mining is a knowledge discovery technique that provides computational intelligence [12]. Text mining techniques are applied in various

ways include categorization of text, summarization, topic detection, concept extraction, search and retrieval, document clustering, etc. Text mining can also be employed to detect a document's main topic/theme which is useful in creating a taxonomy from the document collection [14].

This study focuses on extracting information about how to meet data needs with feature extraction techniques used to enter the text mining process. explicitly the object of the focus outline is described. Author have outline three specific objectives for this SLR

a. To identify the approaches for extraction features from Text Mining Process.
b. To identify tools or libraries used in feature extraction for text mining requirements.

To produce an SRS that is used to fulfill the intended learning, the author divides the explanation of learning outcomes into sub-discussion sections. in section 1, the authors introduce the various introductory framework then the goal of the study. Section 2 is a description of conclusions from related work. The researcher reviews several SLRs that have been made by other researchers and concludes the results of the research. Section 3 is a description of the method carried out in the construction of this SLR. Research questions and literature search methods are explained in this section. Section 4 describes the results of the research conducted on this SLR. in this section, the results of the literature search process are written. then explained Question answer from the research question that was asked. Section 5 provides a discussion of open issues and research implications, and lastly Section 6 provides the concluding remarks.

2. RELATED WORK

While developing this review, researchers also reviewed related products Software Line (SPL), Requirement Reuse, and Text Mining. this section explains summary conclusions from studies related to the object of this paper review.

2.1. Feature extraction approaches from natural language requirements for reuse in software product lines: A systematic literature review [8]

(N. H. Bakar, Z. M. Kasirun, and N. Salleh, 2015) conducted a review that covers a total of 13 primary studies from searching the literature through three main phases: automated database search, complimentary citation-based search, and manual target search. The paper has outlined inclusion and exclusion criteria for selecting the primary studies, which were meant to answer it's main research questions.

The study can supply important contribution to the practitioners and researchers as it provides them with useful information about the different aspects of RR approaches. For practitioners, our SLR has categorised the process for features extraction from NL requirements into phases with detailed information on what approaches are available for adoption in each phase, including some information on tools that are available from open sources.

For researchers, the lower number of selected studies in this SLR indirectly indicates that a lot of research work need to be done in this area. The

2.2. Text mining for market prediction: A systematic review[15]

(A. Khadjeh Nassirtoussi, S. Aghabozorgi, T. Ying Wah, 2014) conclude The major systems for market prediction based on online text mining have been reviewed and some of the predominant gaps that exist within them have been identified. The review was conducted on three major aspects, namely: pre-processing, machine learning and the evaluation mechanism; with each breaking down into multiple sub-discussions. This work intended to accomplish: Firstly, facilitation of integration of research activities from different fields on the topic of market prediction based on online text mining; Secondly, provision of a study-framework to isolate the problem or different aspects of it in order to clarify the path for further improvement; Thirdly, submission of directional and theoretical suggestions for future research.

3. REVIEW METHOD

This section describes the process carried out on this SLR. According to Noor Hasrina Bakar (2015) who cited the opinion of Kitchenham and Charters (2007) that Systematic Literature Review (SLR) is a process of identifying, evaluating, and interpreting all research evidence that has been done to answer certain research questions. SLRs provide a more systematic way of synthesizing evidence of studies that have been conducted specifically with inclusion and exclusion criteria to regulate the boundaries of the evidence to be included in the review paper.

In this SLR paper, the steps of the method used to conduct a review are applied. First, researcher describe the research question that will be discussed. The research question described is related to the extraction technique of text features as the requirement for text-mining requirements to reuse the software product line. the second step is identifying relevant literature and search strategies. The identification of relevant literature is a necessary step in designing a Systematic Literature Review (SLR) [9].

3.1. Identify research questions

In the development of Systematic Literature Review, research questions are used to systematically review a topic [13]. We describe several research questions that relate to the requirements of text-mining for reuse in software product lines.

1) What approaches are available to extract features technique from text mining requirements in the context of software product lines?
2) Are there tools or libraries used in feature extraction for text mining requirements?

3.2. Identification of relevant literature and Search Strategies

Reference search strategies are carried out in several ways. The literature search was carried out using literature references that were used as a reference for making this SLR paper and searched for in the electronic database of journals. The related

literature recommendations from the electronic database of literature search journals are also used as reference materials used. Another search strategy that is carried out is through searching with queries entered in the electronic database of journals and electronic databases from the conference. The electronic literature databases used include:

- IEEE Xplore
- ScienceDirect – Elsevier

Based on the guidelines conducted by Kitchenham and Charters (2007), identification of relevant literature can be done by compiling a search strategy. The initial search can be done on the online literature database. However, there are some challenges to normal online database search: especially the different interface properties for different databases make it difficult to use standard search strings [8]. Therefore, making a complementary manual citation-based (snowballing) search is necessary (Wohlin and Prikladnicki, 2013) to minimise the possibility of missing important evidence.

The literature search process on this SLR consists of three phases following the SLR made by Noor Hasrina Bakar (2015), namely: phase 1: online database search, phase 2: Complementary citation-based search, phase 3: manual target search.

Phase 1: Online database search

Literature search process on this SLR have used the Boolean OR to incorporate synonyms and alternative words. The Boolean AND was used to link the major terms from population, intervention, and context. search queries written in the online search literature database as follows;

```
(("feature extraction" OR "text
mining" OR "requirement") AND
("feature extraction" OR "text
mining" OR "software product
line")AND("feature" OR "requirement"
OR "reuse" OR "software product line"
OR "Systematic Literature
Review")AND("text" AND "feature" AND
"Method"))
```

The literature search on the SLR was conducted in the online literature database written above. In the search process, the authors apply filters to search results. the filter is done, namely, the literature sought is literature with the year of publication between 2018 and the year of making this SLR. then, the author applies the Inclusion and Exclusion criteria and removes irrelevant studies based on title and abstract screening. When titles and abstracts are not enough to identify the relevance of the paper, the full text is then referred to.

Phase 2: Complementary citation-based search

In Phase 2, the author carries out citation searches used in the literature obtained in Phase 1. the author reviews the references used in the selected literature and records relevant and relevant titles as a reference for our SLR. In addition, the authors also look for papers that refer to the selected literature. this can be done using the function of google scholar.

Phase 3: Manual target search

Manually targeted target searches have been proven to bring high-quality search results when combined with the use of searches from digital libraries[6]. The author has included manual targeted searches of the most relevant places I have access to in the areas of Software Engineering and Engineering Requirements in our article search process. Among the leading journals that are used and owned by access are: IEEE Transactions on Software Engineering, IEEE Software, IEEE Systems Journal. This journal is known and used as a reference source for other SLRs related to this SLR [1]. We have searched for all papers published in selected places from January 2008 to December 2018.

4. RESULT

In this section, the author presents the synthesis of evidence from this SLR. the author begins with the analysis of the results from article searches and research present the answers to the main author questions.

4.1. Result of article searches

The literature search process as a reference for the construction of SLR is carried out according to the method described in section 3. Figure 1 is a description of the literature search steps used and the amount of literature obtained.

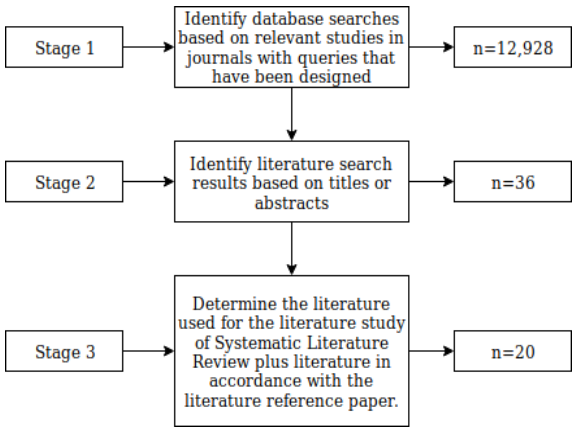

Fig 1. The results of the search process based on the search step

Stage 1 the literature is searched based on the queries that have been designed. the query is entered in the literature search database. literature is obtained in accordance with the number that is drawn, after which it is eliminated based on the appropriate title and abstract in the Stage 2. The title and abstract in the second part of the step are the parameters used to eliminate all the results of the literature search on stage 1. The elimination is carried out by a number of 5 web pages from the results of the literature that appears in its entirety. then stage 3 determination of the literature used for this SLR designer. Literature is generated based on the results of the content selection of the results in stage 2. In addition, the literature from selected literature references is also used in this SLR. The final number of references used is even 20 literature.

4.2. Answering the Research Question

The overall objective of this study was to review the state of current research in the field of feature extraction from text mining requirements for reuse in the SPL. Data

sources that can be used in the text mining process can be obtained from various electronic text data sources. Electronic data sources can be like news data, messages, social media timeline data. Data processing for the social media period is an interesting thing to do with the text mining process. With these data can be found a variety of information that is very useful for many parties.Data sources that can be used in the text mining process can be obtained from various electronic text data sources. Electronic data sources can be like news data, messages, social media timeline data. Data processing for the social media period is an interesting thing to do with the text mining process. With these data can be found a variety of information that is very useful for many parties.

Before entering into the text mining process, the data text must enter the search process for the features that will be processed. There are various types of techniques for extracting text data into new data that can retrieve information from it. To facilitate the elaboration of the explanation, the learning results will be explained based on the response question that was asked. the important sub-explanations to be included are the answers to the related questions.

What approaches are available to extract features technique from text mining requirements in the context of software product lines?

Text mining is a process of how in a data text data mining is carried out so that new information is obtained in it. There are steps that are taken to get information. Figure 2 is a general representation of how the process of text mining[17].

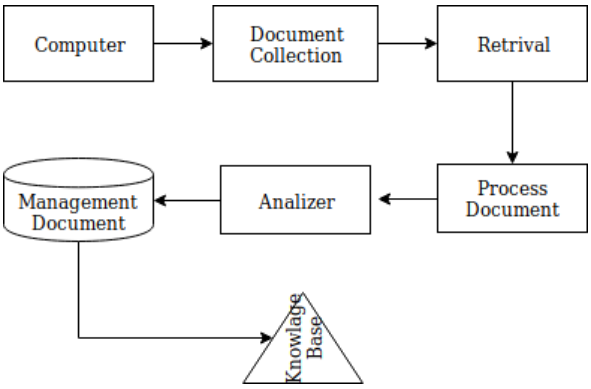

Fig 2. The process is generally text mining

What needs to be underlined in the text-mining process is processing data collection text which can also be said as the preprocessing stage. Data results from preprocessing will be as feature requirements needed in the mining text process. Data processing can help valuable business data from text-based content such as blogs, e-mail, posts, and social media. Intelligent Text Analysis is also called as Text mining. Extracting interesting or interesting useful metadata based text information. The difference between text mining and data mining is that data mining is used to process structured data and metadata based text mining is used to process the unstructured data[18].

Preprocessing methods are commonly carried out, such as Tokenization, Word Removal and Stemming for text documents[19]. After extra preprocessing, the data features that have been repaired can be done. Text feature extraction is a process of how data text is represented as new data, and is generally in the form of a

vector. There are 2 division of approach representations from text, namely Traditional Approach and Word2Vec. A traditional way of representing words is one-hot vector [20]. It is essentially a vector with only one target element being 1 and the others being 0. The word2vec is an open source learning tool based on deep learning, which can convert word into word vector [21]. Two types of Word2Vec is Skip-gram and Continuous Bag of Words (CBOW).

Are there tools or libraries used in feature extraction for text mining requirements?

In producing a data feature, at this time many tools or libraries have been used. The traditional approach and word2vec are provided in the python language library to represent text data into data features. In its current development, the conversion of data text into the word-embedding feature is the most widely used technique. One of the tools used is Glove. GloVe is an unsupervised learning algorithm for obtaining vector representations for words. The training is performed on aggregated global word co-occurrence statistics from a corpus, and the resulting representations of showcase interesting linear substructures of the word vector space[16]. Table 1 presents the tools or libraries that are used according to the existing approach.

Table 1. Approach to extracting features and tools or libraries

Technique	Tool/ Libraries
Traditional Approach : one-hot vector	sklearn.preprocessing. OneHotEncoder
Word2Vec: WordEmbedding	Keras, Tensorflow, Glove

5. DISCUSSION

This section firstly presents a discussion on the implications of this study. The discussion that was opened was how feature extraction data that had been created and produced could support reuse activities in the software product line. The application of text mining in supporting the requirements of reuse in product line software is divided into 4 phases. this phase is adapted to the model of the paper being reviewed [8]. Figure 3 represents a feature extraction process for requirements reuse. Phase 1: assessing requirements, Phase 2: terms extraction (Preprocessing, Word to vector), Phase 3: features identifications and Phase 4: formation of feature models.

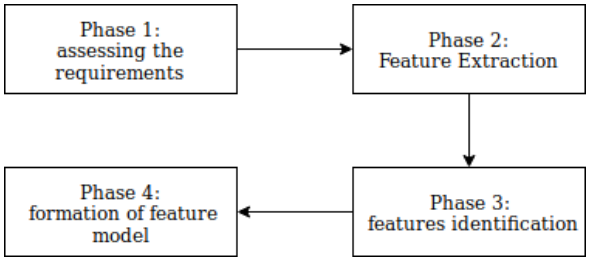

Fig 3. feature extraction process for requirements reuse

6. CONCLUSION AND FUTURE WORK

The author has outlined several techniques about feature extraction carried out in the text. This SLR has answered the research question that was submitted. word to vector or now known as word embedding is a widely developed extraction technique. To support the Reuse in Software Product LIne process there are phases that need to be passed in accordance with the discussion section.

There are many discussions that need to be reviewed in the future. like how the extraction of the features of the text-mining technique is done. In addition, the application of how text mining data to support the reuse process in product line software needs to be discussed in depth. What's more in how the phase so that the results of data extraction can really help.

REFERENCES

[1] Alves, V., Niu, N., Alves, C., Valença, G., 2010. Requirements engineering for software product lines: a systematic literature review. Inf. Softw. Technol. 52 (8)), 806–820.

[2] J.P.T. Higgins, S. Green (Eds.), Cochrane Handbook for Systematic Reviews of Interventions, Version 5.0.0 (updated February 2008), The Cochrane Collaboration, 2008. Available from: .

[3] B.A. Kitchenham and S. Charters, Procedures for performing systematic literature reviews in software engineering: EBSE Technical Report version 2.3, EBSE-2007-01. Keele, UK, 2007.

[4] Barreto, F., Benitti, V., Cezario, R., 2013. Evaluation of a Systematic Approach to Requirements Reuse. J. Univ. Comput. Sci. 19 (2), 254–280.

[5] Benavides, D., Segura, S., Ruiz-Cortes, A., 2010. Automated analysis of feature models 20 years later: a literature review. Inf. Syst. 35, 615–636.

[6] Jørgensen, M., Shepperd, M., 2007. A systematic review of software development cost estimation studies. IEEE Trans. Softw. Eng. 33 (1), 33–53.

[7] Wohlin, C., Prikladnicki, R., 2013. Systematic literature reviews in software engineering. Inf. Softw. Technol. 55 (6), 919–920.

[8] N. H. Bakar, Z. M. Kasirun, and N. Salleh, "Feature extraction approaches from natural language requirements for reuse in software product lines: A systematic literature review," *J. Syst. Softw.*, vol. 106, pp. 132–149, 2015.

[9] T. Dybå and T. Dingsøyr, "Empirical studies of agile software development: A systematic review," *Inf. Softw. Technol.*, vol. 50, no. 9–10, pp. 833–859, 2008.

[10] M. Marques, J. Simmonds, P. O. Rossel, and M. C. Bastarrica, "Software product line evolution: A systematic literature review," Inf. Softw. Technol., vol. 105, pp. 190–208, Jan. 2019.

[11] L. M. Northrop et al., "A FRAMEWORK FOR SOFTWARE PRODUCT LINE PRACTICE,

VERSION 5.0 What's New in Version 5.0," 2012.

[12] A. Rosenfeld, M. E. Taylor, and S. Kraus, "Leveraging human knowledge in tabular Reinforcement learning: A study of human subjects," IJCAI Int. Jt. Conf. Artif. Intell., vol. 00, pp. 3823–3830, 2017.

[13] B.A. Kitchenham and S. Charters, Procedures for performing systematic literature reviews in software engineering: EBSE Technical Report version 2.3, EBSE-2007-01. Keele, UK, 2007.

[14] H. Hashimi and A. Hafez, "Selection criteria for text mining approaches," Comput. Human Behav., vol. 51, pp. 729–733, Oct. 2015.

[15] A. Khadjeh Nassirtoussi, S. Aghabozorgi, T. Ying Wah, and D. C. L. Ngo, "Text mining for market prediction: A systematic review," Expert Syst. Appl., vol. 41, no. 16, pp. 7653–7670, Nov. 2014.

[16] C. Mackay, "'Glove.,'" Notes Queries, vol. s5–IV, no. 96, p. 346, 1875.

[17] S. S. Bhanuse, S. D. Kamble, and S. M. Kakde, "Text Mining Using Metadata for Generation of Side Information," Procedia Comput. Sci., vol. 78, pp. 807–814, Jan. 2016.

[18] T. Yang, R. Jin, Y. Chi, and S. Zhu, "Combining link and content for community detection: A discriminative approach," in Proc.ACM KDD Conf., New York, NY, USA, 2009, pp. 927–936.

[19] C. Paper, "Preprocessing Techniques for Text Mining Preprocessing Techniques for Text Mining," J. Emerg. Technol. Web Intell., no. October 2014, 2016.

[20] A. Hassan and A. Mahmood, "Convolutional Recurrent Deep Learning Model for Sentence Classification," IEEE Access, vol. 6, pp. 13949–13957, 2018.

[21] [1] Zhi-Tong Yang and Jun Zheng, "Research on Chinese text classification based on Word2vec," in 2016 2nd IEEE International Conference on Computer and Communications (ICCC), 2016, pp. 1166–1170.

[22] "Word2Vec and FastText Word Embedding with Gensim – Towards Data Science." [Online]. Available: https://towardsdatascience.com/word-embedding-with-word2vec-and-fasttext-a209c1d3e12c. [Accessed: 11-Dec-2018