Requirement Formalisation Using Natural Language Processing and Machine Learning: A Systematic Review

Shekoufeh Kolahdouz-Rahimi[1][a], Kevin Lano[2][b] and Chenghua Lin[3][c]
[1]*School of Arts, University of Roehampton, London, U.K.*
[2]*Department of Informatics, King's College London, London, U.K.*
[3]*Department of Computer Science, University of Sheffield, U.K.*
fi

Keywords: Requirements Engineering, Requirement Formalisation, Natural Language Processing, Machine Learning, Deep Learning, Systematic Mapping Study.

Abstract: Improvement of software development methodologies attracts developers to automatic Requirement Formalisation (RF) in the Requirement Engineering (RE) field. The potential advantages of applying Natural Language Processing (NLP) and Machine Learning (ML) in reducing the ambiguity and incompleteness of requirements written in natural languages are reported in different studies. The goal of this paper is to survey and classify existing works on NLP and ML for RF, identifying the challenges in this domain and providing promising future research directions. To achieve this, we conducted a systematic literature review to outline the current state-of-the-art of NLP and ML techniques in RF by selecting 257 papers from commonly used libraries. The search result is filtered by defining inclusion and exclusion criteria and 47 relevant studies between 2012 and 2022 are selected. We found that heuristic NLP approaches are the most common NLP techniques used for automatic RF, primarily operating on structured and semi-structured data. This study also revealed that Deep Learning (DL) techniques are not widely used, instead, classical ML techniques are predominant in the surveyed studies. More importantly, we identified the difficulty of comparing the performance of different approaches due to the lack of standard benchmark cases for RF.

1 INTRODUCTION

Productive management of Requirement Engineering (RE) accelerates the process of software development. Requirement Formalisation (RF) relates to the process of transforming requirements in natural language to specific formal notations by removing ambiguities. Formal specification of requirements is applicable in different stages of software development especially in the validation phase. Manual formalisation of natural language requirements is an error-prone and time-consuming task and infeasible for complex systems (Zaki-Ismail et al., 2021). To this end many automatic and semi-automatic approaches have been introduced to formalise requirements by applying Natural Language Processing (NLP) techniques (Rolland and Proix, 1992; Ryan, 1993). Additionally, leveraging Machine Learning (ML) and Deep Learning (DL) techniques in this domain pushes the research field forward.

In the last decade, there has been a noticeable increase in the number of papers using NLP and ML techniques for RF. Each research applied a particular technique and mostly there are no comprehensive guidelines for the reason of applying those techniques. A large number of research reviews have been carried out to survey this domain (Alzayed and Al-Hunaiyyan, 2021). To overcome the limitations of existing studies, in this paper we conducted a systematic mapping study of NLP and ML approaches for RF considering the guidelines presented by Kitchenham and Charters (Brereton et al., 2007), (Keele et al., 2007), and Petersen et al. (Petersen et al., 2015). We investigated 47 studies from an initial set of 257. The papers are selected from commonly used libraries including ACM Digital Library, IEEE Xplore, ScienceDirect, Springer Link, and Scopus. The search results are filtered by defining inclusion and exclusion criteria to decide whether a publication found in the search should be included in the study or excluded. Three research questions were formulated for our Systematic Literature Review (SLR). By answering these research questions, the state of the art of RF

[a] https://orcid.org/0000-0002-0566-5429
[b] https://orcid.org/0000-0002-9706-1410
[c] https://orcid.org/0000-0003-3454-2468

Kolahdouz-Rahimi, S., Lano, K. and Lin, C.
Requirement Formalisation Using Natural Language Processing and Machine Learning: A Systematic Review.
DOI: 10.5220/0011789700003402
In *Proceedings of the 11th International Conference on Model-Based Software and Systems Engineering (MODELSWARD 2023)*, pages 237-244
ISBN: 978-989-758-633-0; ISSN: 2184-4348

is evaluated. We identify current challenges in the community and provide guidelines to address those challenges. The actions that were taken by the author of this paper as part of MDENet project [1] are also discussed. Finally, for further maturity of the research field, future research directions in this area are provided.

The remainder of this paper is organised as follows: In Section 2, related review papers in RF using NLP and ML are summarised. Section 3, presents background information related to this domain. The applied research method in this study is described in Section 4. Section 5 outlines the key findings of our study. A discussion of action taken by the authors to address the challenges is provided in Section 6. Finally, the conclusion and future direction of research are provided in Sections 7 and 8.

2 RELATED WORK

A comprehensive survey in the application of NLP to RE is provided in (Zhao et al., 2020), by considering 404 works. This paper emphasise is on the insufficient application of NLP techniques for RE studies in industrial cases. Importantly, the lack of expertise in selecting appropriate NLP techniques in RE domain is discussed. Although challenges are introduced in this work, practical solutions are not provided. Additionally, investigating ML techniques in this domain is not the main focus of research.

A survey in the application of NLP techniques to requirements in the form of user stories is provided in (Raharjana et al., 2021) and the potential advantages of those techniques in RF domain are discussed. A comprehensive classification based on the uses of NLP techniques for user stories is provided and a model is recognised as a common target for RF in the form of user stories. However, ML techniques for RF are not considered in this research, and in general, the main challenges of the domain are not sufficiently discussed. Additionally, input datasets for user stories in investigated papers are not provided.

Yalla and Sharma (Yalla and Sharma, 2015) survey the current literature that leverages RE and NLP for different phases of software development. However, only limited future research directions and guidelines are provided in this work. Selected articles that generate UML diagrams by applying NLP techniques are investigated in (Dawood et al., 2017; Abdelnabi et al., 2021). These works emphasize the immaturity of the research area as most of the current processes are not automated. The advantages and disadvantages of different studies that generate UML diagrams by applying heuristic rules are identified in (Ahmed et al., 2022). This work highlighted the noticeable application of ML in this domain.

The are many related literature studies in the RF domain. However, the challenges of the domain are not deeply recognized and guidelines for addressing those challenges and clear research direction are limited in most studies. This proves the immaturity of the research area and its potential for further improvement. Therefore, the main aim of this research is to identify current challenges in the community by classifying applied techniques and providing practical guidelines for addressing those issues.

3 BACKGROUND

In this section, the related concepts to this research including RE, NLP and DL are explained.

3.0.1 Requirement Formalisation

RE is an important process in software development for discovering stakeholder's needs and classifying them for other phases of software development (Pohl, 2010). Formalising requirement is one of the key tasks in RE that is automated by applying different NLP and ML techniques and tools. It enables the translation of the requirement in natural language into a structured formal form (e.g., UML modeling diagrams). This transformation reduces the ambiguities of natural language and provides a convenient way for validation and verification (Schön et al., 2017; Tukur et al., 2021).

3.1 Natural Language Processing

NLP is an area of research in Artificial Intelligence (AI) that enables a computer to process a large amount of structured/unstructured data in natural language that exist in today's world. Different NLP methods, approaches, processes, and procedures are introduced to process data in different phases of RE. Tokenization, POS tagging, and dependency parsing are the commonly used techniques in this domain (Kulkarni and Shivananda, 2019).

3.2 Machine Learning

ML is one of the core technical terms in Artificial Intelligence (AI), which refers to the learning and identification of patterns from examples and existing data

[1] https://mde-network.com/

Table 1: Terms for Selecting Relevant Research Studies.

Group	Term
A	Natural Language Processing
	Natural Language, NLP
B	Machine Learning
	Deep Learning
C	Requirement Formalisation
	Model Generation
	UML Generation
	OCL Generation
	Usecase or Use case Diagram Generation
	Class Diagram Generation
	Sequence Diagram Generation
	ER Diagram Generation
	Activity Diagram Generation

(Samuel, 1967). Different algorithms and processing techniques are introduced in this domain. DL is an ML technique, which is based on Artificial Neural Network (ANN) (Goodfellow et al., 2016) and is applicable in a variety of domains including RF.

4 RESEARCH METHOD

The provided guidelines by Kitchenham and Charters (Brereton et al., 2007), (Keele et al., 2007), and Petersen et al. (Petersen et al., 2015) are applied for systematic mapping study in this research. The following research questions are the main target of this research.

- **Q1:** What are the most commonly used NLP/ML approaches for automatic/semi-automatic RF?
- **Q2:** What are the input and output of RF approaches?
- **Q3:** What are the gaps and deficiencies in existing RF work?

The three phases of the study protocol including planning, conducting, and reporting are explained in the following sections.

4.1 Review Planning

The review process and search strategy are explained in this part. To provide comprehensive coverage of existing publications most major publishers in Software Engineering are investigated.

According to the objectives of this study and research questions, three terms were selected in this paper. Each term includes different keywords and at least one of the keywords has to be presented in a paper. Table 1 presents the list of selected terms in this research.

To identify the largest number of studies in the domain of RF, the following search string is followed:

Search String= $(A \vee B \vee (A \wedge B)) \wedge C$

Additionally, inclusion and exclusion criteria to decide which of the selected articles should be considered as primary studies and which ones should be excluded are defined as follows:

4.1.1 Inclusion Criteria

- Published between January 2012 and March 2022
- Publications that generate model or any formalisation from requirement
- Publications in peer-reviewed journals, conferences, and workshops
- Publication in English

4.1.2 Exclusion Criteria

- Publications not written in English
- Publications before 2012
- Summary, survey, or review publications
- Non peer-reviewed publications
- Publications not focusing on RF
- Books, web sites, technical reports, pamphlets, tutorials, duplicate papers, and white papers.

In this research abstracts, titles, and keywords of papers are evaluated according to the inclusion and exclusion criteria. Furthermore, in some cases the whole text of the paper is also investigated.

4.2 Review Conducting

The review conduction stage presents the selection process for LR in this research.

4.2.1 Article Selection

This phase is divided into three sub tasks including, pilot study, article selection and quality assessment of selected primary studies.

Pilot Study. A pilot study is carried out to investigate the reliability of provided selection criteria as suggested by Kitchenham and Charters (Brereton et al., 2007), (Keele et al., 2007), and Petersen et al. (Petersen et al., 2015) before the selection of primary articles. In this stage, five papers are selected by the first and second authors. These articles are investigated by a third author, who is an expert in the NLP domain and was not involved in the search process by considering the inclusion and exclusion criteria. A satisfactory result is presented from the pilot study, which proves the suitability of the defined criteria in this research.

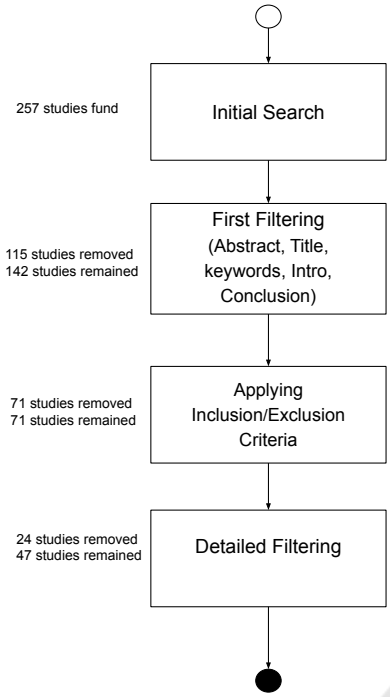

Figure 1: Primary studies selection process.

Primary Study Selection. In this part, the relevant articles are searched using the provided search string. The result of this selection is presented in Figure 1. For the initial search process, 257 results are recognised. To refine the selected papers in the next iteration, titles, abstracts, keywords, introduction, and conclusion sections are reviewed. As a result, 115 papers are removed from the list of selections, and then the rest are kept for the next iteration. Following that by applying inclusion and exclusion criteria 71 papers are rejected. Next for this iteration, the content of the paper is investigated deeply and 24 papers are rejected. Finally, 47 studies are remaining. Figure 3 presents the distribution of the resulting papers in each year. As can be seen in this figure, in 2021 the domain gained more interest and the highest number of publications were published. Additionally, Figure 2 indicates the publication type of result papers. Conferences are the target of publication in most studies.

Quality Assessment. The quality of the selected study is also assessed in this research. Therefore, a checklist with four quality assessment questions is presented in Table 2. The questions are answered by the first author by selecting from 'yes', 'no', and 'partly' options.

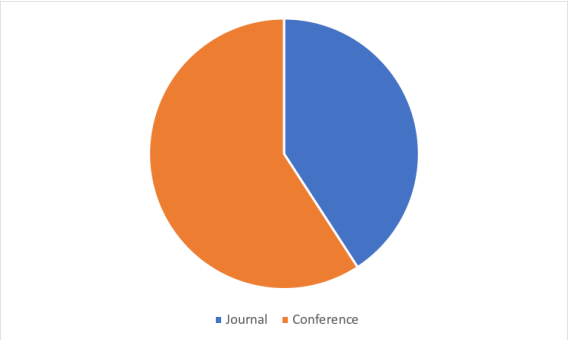

Figure 2: Primary studies per publication type.

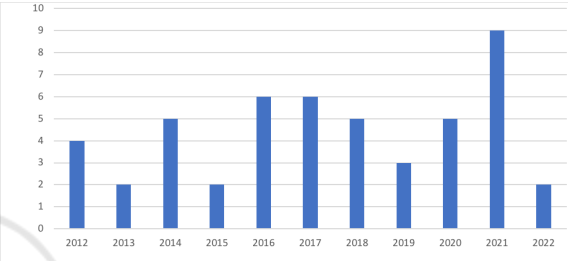

Figure 3: Distribution of NLP papers in each year.

4.2.2 Data Extraction and Synthesis

To answer each research question, the data extraction process is performed by developing a predefined data extraction form in Table 3. The form enables us to record essential information about primary studies to answer each research question. The form is filled out by the first author manually and then the second and third authors reviewed the results and finally, the issues are fixed.

4.2.3 Reporting the Review

Based on the results of data extraction phase, the review result is presented and each research question is answered and discussed.

Table 2: Quality assessment Questions.

QID	Topic	Question
A1	Objective	Did the study clearly define the research objectives?
A2	Related work	Did the study provide a review of previous work?
A3	Research methodology	Are the research methodology clearly established?
A4	Validity	Did the study include a discussion on the validity and reliability of the procedure used?
A5	Future work	Did the study point out potential further research?

Table 3: Data Extraction Form.

Study data	Description	Relevant RQ
Title		Study Overview
Author		Study Overview
Year		Study Overview
Article Source		Study Overview
Type of Article	Journal, Conference Workshop	
Research goal	What is the main goal of study?	RQ1
Research goal category	Model extraction/generation, Requirement formalisation, UML, Usecase, Class Activity, ER diagram extraction/generation	RQ1
Research method	What research methods did the study employ?	RQ1
Data	What are the datasets for evaluation of the study	RQ2
Evaluation	what are the evaluation criteria in the study?	RQ1
NLP techniques	What are the applied NLP techniques?	RQ1
ML techniques	What are the applied ML techniques?	RQ1
NLP tools	What NLP tools did the study use?	RQ1
Challenges	What are the challenges of the study?	RQ3
Future Work	What are the suggested future work?	RQ3

5 RESULTS

This section presents the result of the review of this research. We selected 47 primary studies for SLR.

5.1 Summary of Studies

Investigated studies for classifying NLP and ML techniques to formalise requirements are available in (RFr, 2022). The studies are summarised according to the applied NLP/ML techniques, input and output artifacts, datasets, approach, applied tools and libraries, input structure, and evaluation criteria.

This section discusses the classification results of the investigated approaches according to each research question.

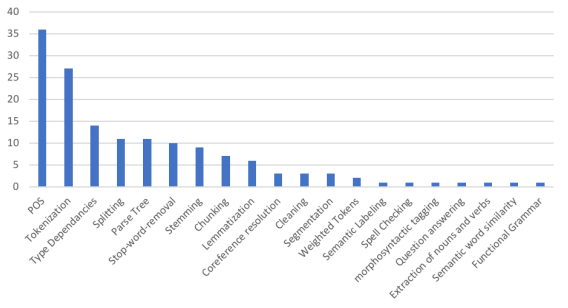

Figure 4: Frequency of NLP Technologies in Different Studies.

5.2 Q1: What Are the Most Commonly Used NLP/ML Approaches for Automatic/Semi-Automatic RF?

To answer this question, NLP and ML approaches applied in selected studies are deeply investigated.

5.2.1 NLP Techniques

Figure 4 presents the applied NLP techniques and frequency of using those techniques through out investigated papers.

- **Applied Techniques.** Tokenization, POS tagging and Type dependencies are the most common used NLP techniques in those research.
- **Heuristic Rules.** Majority of research applied heuristic rules for formalising requirements.
- **Frequent Tools and Libraries.** Stanford core NLP is the most common-used NLP tools in investigated studies.
- **Evaluation Criteria.** Accuracy in terms of precision, recall and F-measure are the most common criteria for evaluation in selected studies.

5.2.2 ML Techniques

There are not many studies that apply ML techniques for RF and mostly classical ML techniques such as decision trees and Support Vector Machine (SVM) are used in those studies. The applied ML techniques and frequency of application of these techniques are presented in Figure 5. Around 20% of selected studies in this research applied both NLP and DL techniques.

5.3 Q2: What Are the Input and Output of RF Approaches?

This question is answered according to the applied type and structure of input elements and generated output elements.

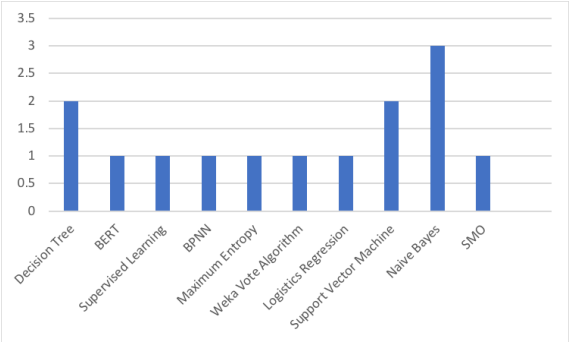

Figure 5: Frequency of DL Technologies in Different Studies.

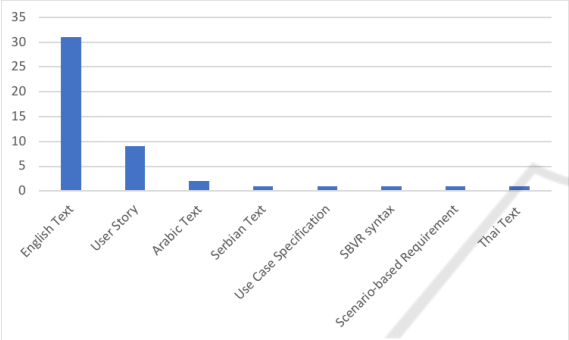

Figure 6: Frequency of Input for RF.

- **Input Types.** The frequency of input types in selected studies is presented in Figure 6. English text and user story are the most common-used type in most studies.
- **Input Structure.** The unstructured English text is the most frequent input structure for most of the studies as presented in Figure 7. The inputs in the format of user story are semi-structured.
- **Output Types.** In most studies, the formalisation is in the form of UML diagrams. Figure 8 indicates that class and use case diagrams are the most common input types in these studies.

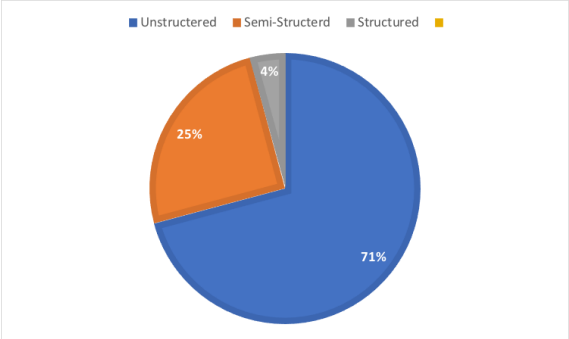

Figure 7: Structure of Input Elements.

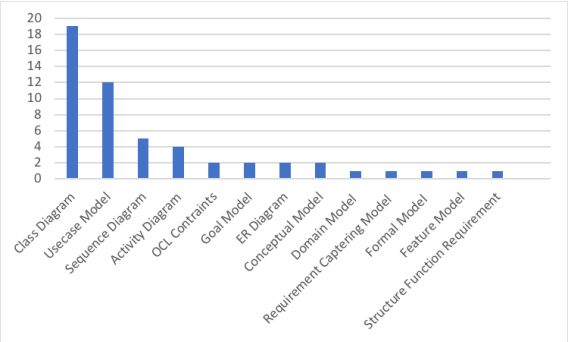

Figure 8: Frequency of Generated Output for RF.

5.4 Q3: What Are the Gaps and Deficiencies in Existing RF Work?

The RF field remains at an experimental stage, in particular evaluation of approaches is not performed systematically and it is difficult to compare different approaches. The published results of studies were often not reproducible due to the unavailability of tools or data.

Investigating different works, we identified three main deficiencies for formalising requirements. A list of deficiencies is provided below:

5.4.1 Lack of Completeness in Heuristic Rules

The performance and completeness of heuristic approaches are typically not evaluated on a broad range of input cases, thus it is not possible to determine which are the best to use in different situations. A lot of papers applied heuristic approaches and defined rules manually. It is not possible to come up with any specific number of rules for formalising requirements and generating relevant artifacts. This issue is not deeply investigated in the community as there is not an adequate comparative evaluation in this domain.

5.4.2 Lack of Application of DL

There is under-use of DL techniques, which seem to be relevant and applicable to RF tasks and could help to avoid the limitations of heuristic approaches, especially for unstructured source data. It is assumed that the limited number of training data in the community is the main reason that developers do not use DL models in different tasks. Most of the learning methods used in the community are standard methods such as decision tree or regression model. It can be concluded that most of the studies use learning in the wrong way and do not exploit the full potential of deep learning techniques in this domain. Therefore, in theory, application of ML provides a potential solution for the

different task of RF. Community can benefit from the DL models and tools by applying some modern learning architecture such as OpenAI Codex (Chen et al., 2021) as in these models it is not essential to have lots of labels data for performing particular task.

5.4.3 Lack of Evaluation Benchmark Framework

To systematically compare different RF cases, standard benchmarks and evaluation criteria need to be established e.g., there are well-established benchmarks in Natural Language Generation (GEM,) and Speech Processing (SUP,). Many of the evaluation datasets cited in selected papers are no longer available. Therefore, a repository of standard cases, proposed approaches, and evaluation procedures are necessary. This is an important issue in the community and it is essential to fill this gap.

6 DISCUSSION ON ACTION TAKEN

Our systematic review results show open issues and research challenges for formalising requirements. This research is part of MDENet project and some actions are taken by the authors of this paper to solve part of the issues. These actions are summerised below:

- In order to strengthen the area of RF research, we developed a DSL for NLP pipelines, based on the SQLite grammar of GitHub - antlr/grammars-v4/SQLite. This enables the high-level definition of NLP pipelines for RF, independent of any particular NLP platform such as NLTK or Apache OpenNLP. Common RF processing such as POS-tagging, segmentation, chunking, and parsing can be specified. A transformation from the DSL to Python was defined to support implementation in NLTK.
- To provide a central point of reference for RF research, we established a GitHub repository (RFr, 2022), which will hold links to state-of-the-art research in the area, evaluation cases, evaluation tools, and the results of evaluations. The repository will be a resource for the RF community and aims to improve the practical application of RF research to real-world software problems.
- To compare the effectiveness of RF approaches, there needs to be an established set of requirements statements that can be applied. We selected 25 cases of real-world requirements statements in the format of user stories, which reflect a diversity of linguistic styles and scales, and added these to the RF repository.
- To evaluate the results of applying RF approaches to the evaluation cases, we provide tools to (i) compare the formalized models produced by an approach to manually constructed reference models for the cases, to identify a measure of similarity of these models; (ii) compare the formalized models to the source document, to check the completeness of the formalization; (iii) to evaluate the internal quality of the formalized model. Three kinds of evaluation strategy are presented in Figure 9. Example evaluations have been provided for three RF approaches, evaluated on two user story case studies.

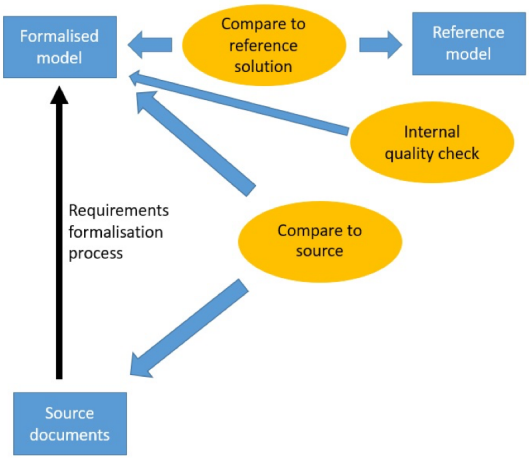

Figure 9: Three Kinds of Evaluation Strategy for RF Approaches.

7 FUTURE DIRECTIONS

In the following we discuss directions to complete the current actions and further future work on RF.

- Generate a platform to guide the user in selecting appropriate NLP techniques occurring to their requirements. This will occur by considering more case studies and evaluating them by applying the evaluating tools specified in (RFg, 2022).
- We will develop the RF repository with more example case studies, including examples of unstructured requirements statements/background documentation, evaluation tools and evaluations, and publicise this in MDE forums and invite contributions from RF researchers.

8 CONCLUSION

This research carried out a systematic survey of existing approaches for RF, including NLP and ML approaches across a wide range of applications. 250 publications were examined, and 47 specific publications were selected for deeper analysis. We identified that:

- Heuristic NLP approaches are the most common RF technique in the research, primarily operating on structured and semi-structured data.
- Deep learning techniques are not widely-used, instead classical ML techniques such as decision trees and Support Vector Machine (SVM) are used in the surveyed studies.
- There is a lack of standard benchmark cases for RF and therefore it is difficult to compare the performance of different approaches.

REFERENCES

Gembenchmark. https://gem-benchmark.com/.

Superbbenchmark. https://superbbenchmark.org/.

(2022). Repository of state-of-the art requirements formalisation approaches.

(2022). Requirement formalisation using natural language processing and machine learning repository. https://doi.org/10.5281/zenodo.7337229/.

Abdelnabi, E. A., Maatuk, A. M., and Hagal, M. (2021). Generating uml class diagram from natural language requirements: A survey of approaches and techniques. In *2021 IEEE 1st International Maghreb Meeting of the Conference on Sciences and Techniques of Automatic Control and Computer Engineering MI-STA*, pages 288–293. IEEE.

Ahmed, S., Ahmed, A., and Eisty, N. U. (2022). Automatic transformation of natural to unified modeling language: A systematic review. *arXiv preprint arXiv:2204.00932*.

Alzayed, A. and Al-Hunaiyyan, A. (2021). A bird's eye view of natural language processing and requirements engineering. *Int. J. Adv. Comput. Sci. Appl.*, 12(5):81–90.

Brereton, P., Kitchenham, B. A., Budgen, D., Turner, M., and Khalil, M. (2007). Lessons from applying the systematic literature review process within the software engineering domain. *Journal of systems and software*, 80(4):571–583.

Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. d. O., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., et al. (2021). Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Dawood, O. S. et al. (2017). From requirements engineering to uml using natural language processing–survey study. *European Journal of Industrial Engineering*, 2(1):pp–44.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep learning*. MIT press.

Keele, S. et al. (2007). Guidelines for performing systematic literature reviews in software engineering. Technical report, Technical report, ver. 2.3 ebse technical report. ebse.

Kulkarni, A. and Shivananda, A. (2019). *Natural language processing recipes*. Springer.

Petersen, K., Vakkalanka, S., and Kuzniarz, L. (2015). Guidelines for conducting systematic mapping studies in software engineering: An update. *Information and software technology*, 64:1–18.

Pohl, K. (2010). *Requirements engineering: fundamentals, principles, and techniques*. Springer Publishing Company, Incorporated.

Raharjana, I. K., Siahaan, D., and Fatichah, C. (2021). User stories and natural language processing: A systematic literature review. *IEEE Access*, 9:53811–53826.

Rolland, C. and Proix, C. (1992). A natural language approach for requirements engineering. In *International conference on advanced information systems engineering*, pages 257–277. Springer.

Ryan, K. (1993). The role of natural language in requirements engineering. In *[1993] Proceedings of the IEEE International Symposium on Requirements Engineering*, pages 240–242. IEEE.

Samuel, A. L. (1967). Some studies in machine learning using the game of checkers. ii—recent progress. *IBM Journal of research and development*, 11(6):601–617.

Schön, E.-M., Winter, D., Escalona, M. J., and Thomaschewski, J. (2017). Key challenges in agile requirements engineering. In *International Conference on Agile Software Development*, pages 37–51. Springer, Cham.

Tukur, M., Umar, S., and Hassine, J. (2021). Requirement engineering challenges: A systematic mapping study on the academic and the industrial perspective. *Arabian Journal for Science and Engineering*, 46(4):3723–3748.

Yalla, P. and Sharma, N. (2015). Integrating natural language processing and software engineering. *International Journal of Software Engineering and Its Applications*, 9(11):127–136.

Zaki-Ismail, A., Osama, M., Abdelrazek, M., Grundy, J., and Ibrahim, A. (2021). Arf: Automatic requirements formalisation tool. In *2021 IEEE 29th International Requirements Engineering Conference (RE)*, pages 440–441. IEEE.

Zhao, L., Alhoshan, W., Ferrari, A., Letsholo, K. J., Ajagbe, M. A., Chioasca, E.-V., and Batista-Navarro, R. T. (2020). Natural language processing (nlp) for requirements engineering: A systematic mapping study. *arXiv preprint arXiv:2004.01099*.