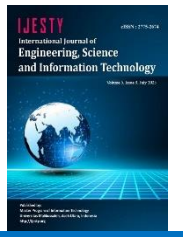
International Journal of Engineering, Science and Information Technology
Volume 5, No. 3 (2025) pp. 480-490
eISSN: 2775-2674
Website: http://ijesty.org/index.php/ijesty
DOI: https://doi.org/10.52088/ijesty.v5i3.1161
Research Paper, Short Communication, Review, Technical Paper



Exploring the Synergy: User Stories in Agile Software Development

Siti Nur Fathin Najwa Binti Mustaffa*, Jamaludin Bin Sallim, Rozlina Binti Mohamed

Faculty of Computing, Universiti Malaysia Pahang Al-Sultan Abdullah, Malaysia

**Corresponding author Email: fatinnajwa22@gmail.com*

The manuscript was received on 10 January 2025, revised on 10 June 2025, and accepted on 24 June 2025, date of publication 9 July 2025

Abstract

User Stories are commonly used artifacts to capture user requirements in Agile Software Development (ASD). They are short, semi-structured statements that describe requirements. Natural Language Processing (NLP) techniques can be advantageous for research on user stories. This paper investigates User Stories and NLP about their applications, critically examines existing research approaches related to NLP in user stories, presents the challenges and suggested future work. Relevant papers were obtained from well-recognized digital libraries and scientific databases, including ScienceDirect, Scopus, SpringerLink, and IEEE Xplore. Inclusion and exclusion criteria were applied to filter search results and obtain comprehensive findings. The search results identified 1175 papers published between 2014 until 2024. After applying the inclusion/exclusion criteria, 35 primary studies discussing NLP techniques in user stories were selected. The purposes of these studies vary, encompassing defect discovery, software artifact generation, key abstraction identification in user stories, and linking models and user stories. NLP can assist system analysts in managing user stories. Implementing NLP in user stories offers numerous opportunities and challenges. Exploring NLP techniques and employing rigorous evaluation methods are necessary for high-quality research. As with general NLP research, understanding the context of sentences remains a challenge.

Keywords: *Agile Software Development, Natural Language Processing, Systematic Review, User Story, Quality.*

1. Introduction

1.1. Agile Software Development

Since the release of the Agile Manifesto [1][2], there has been a notable increase in the adoption of Agile Software Development (ASD) practices. This growing trend is driven by the dynamic nature of today's business environment, which demands adaptable and responsive systems to maintain organizational and operational effectiveness [3][4]. ASD caters to these demands by offering numerous advantages, such as delivering high-quality products [1][2], optimizing resource utilization [3], accelerating software development, and allowing for flexible and evolving requirements [4][5]. Beyond its focus on software design and coding, ASD also emphasizes Requirements Engineering (RE) activities throughout the development lifecycle [3], [6]. Design and coding are essential for fulfilling software requirements effectively [7], while RE plays a crucial role in enhancing the quality of digital services [8][9]. RE activities include requirements elicitation, documentation, analysis, negotiation, validation, and management [2][10]. ASD enables a flexible, iterative approach to handling and updating requirements, even during the later phases of development [2][8]. However, these processes necessitate strong collaboration both within the development team and between developers and users to ensure high software quality, on-time delivery, customer satisfaction, and alignment with product expectations [11][12].

1.2. User Stories

User stories are increasingly gaining a place in the software development process, especially in Agile Software Development (ASD). User Stories are the most widely used artifact in ASD [1][2] that express requirements form the user points of view. A user story is a semi structured specification of requirement written in natural language. A user story template may take the following form [3]: as [WHO], I want/want to/need/can/would like [WHAT], so that [WHY]. It contains important elements of requirements. WHO wants it, WHAT is expected from the system and optionally and WHY it is important [3],[4]. A user story also must be a short, semi-structured sentence that illustrates requirements from the user's perspective and can be used to explain user desire or product description [12]. The aspect of "who" refers to the system user or actor, "what" refers to the actor's desire, and "why" refers to the reason (optional in the user story). These aspects are arranged into one sentence with a certain structure. Several formats or templates are usually used as per Figure 1.

"As a <persona / actor >, I want to < aspect of what / task > so that < aspect of why/ goal >"
OR
"As a < persona/ actor >, I need < aspect of what / task > so that < aspect of why/ goal >"
OR
"As a < persona/ actor >, I can < aspect of what / task > so that < aspect of why/ goal >"

Fig 1. Format of written User Stories

The user story components consist of the following elements [13]:
1. Role: abstract behavior of actors in the system context; the aspect of who representation
2. Goal: a condition or a circumstance desired by stakeholders or actors
3. Task: specific things that must be done and achieve goals
4. Capability: the ability of actors to achieve goals based on certain conditions and events

The rise of ASD has attracted researchers and practitioners into this research field [1][5][6]. User Stories, as the most widely used artifacts in ASD, are challenging to explore. The fact that they are written in natural languages makes them easy to understand to stakeholders. However, requirements written in natural language have drawbacks such as ambiguity, inconsistency, and incompleteness. [7][12]. This paper structure consists of following sections; Section 2 discusses the previous related review, Section 3 presents the research methodology used to conduct this Systematic Literature Review (SLR), Section 4 presents a discussion of the threat validity of this review, Section 5 outlines the detailed descriptions of the results and findings of the review by providing answers for each specific research question and Section 6 elaborates upon discussion regarding the research findings and identifies the study limitations.

2. Literature Review

Studies conducted have focused on the user story's specification to the best of our knowledge. Several secondary studies related to this area focus on several issues, aspects and areas (e.g., Agile Requirements Engineering [1][5], Quality Requirement management in Agile Software Development [14], the evolution of use cases [14], and Requirements Engineering in Model-Driven development [15]). Table 1 summarizes these works.

Table 1. Summarization of existing research

No.	Year	Researcher	Main Focus
1.	2017	Scon et al [1]	Focused on the stakeholders and user involvements
2.	2015	Inayat et al [5]	Focused on adapting agile requirements engineering practices.
3.	2019	Behutiye et al [16]	Focused on the review, quality requirement management and agile software development were not specifically discussed
4.	2018	Heck et al [17]	Focused on the quality criteria for evaluating the correctness of agile requirements
5.	2015	Tiwari and Gupta [14]	Focused on the review's studies related to the evolution of use cases. Use cases are artifacts with almost the same functions as user stories. They stated that use cases increasingly utilize formal structures to facilitate software development life cycle (SDLC) activities. The researcher also performed comparisons of the development of use case and user stories to obtain more appropriate comparison
6.	2017	Loniewski et al [18]	Focused on review studies related to usage user requirements engineering techniques for model driven development. The NL requirements are usually used for the automation of the SDLC process
7.	2015	Bakar et al [20]	Focused on the extracting NL requirements for reuse in software product line engineering
8.	2017	Nazir et al [21]	Focused on the NL application in software requirements

3. Methods

3.1. Review Methods

We adopted procedures from [22] and [23] in preparing the SLR comprising three stages which are review planning, conducting, and reporting. The 2009 PRISMA checklist was adopted as a guide in writing this SLR Report [24]. Figure 2 shows the review protocol which we adopted consists of three (3) phases: Plan Review, Conduct Review and Document Review. Each of the phases consists of different activities which the step is mentioned according to. Starting with specific research questions, search processes involving inclusion and exclusion criteria, identify relevant research for selection of primary studies. Next, the data was extracted and synthesized by applying the quality assessment criteria. Once all data has been collected and synthesized, then we start with the writing report and validate the results.

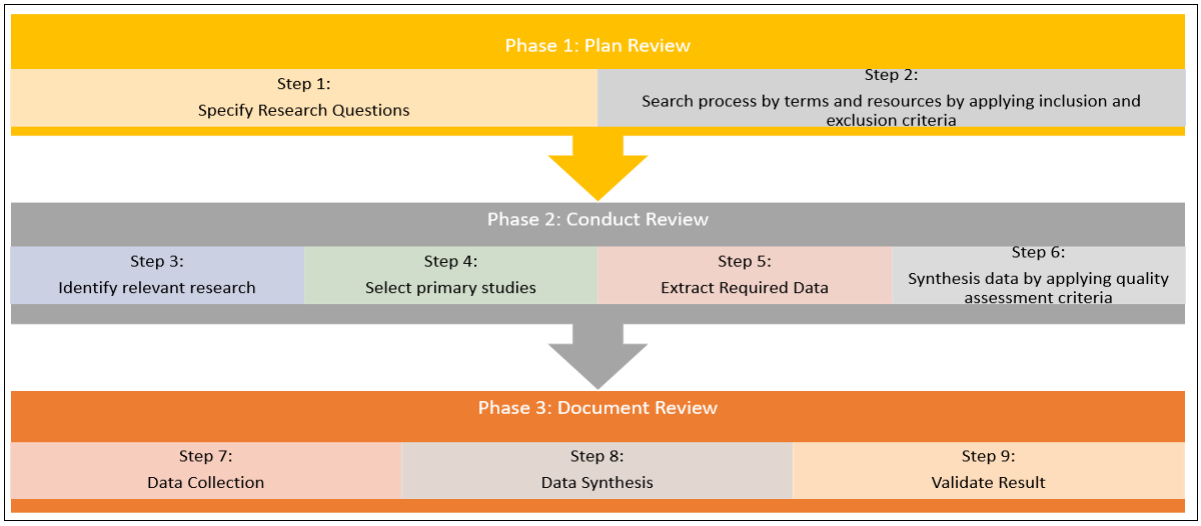


Fig 2. Review protocol

3.2. Research Objectives (RO) and Research Question (RQ)

The rise of Agile Software Development (ASD) research has led to an increase of research related to the user stories, which are the most widely used artifacts in ASD. The user story format that uses natural languages makes the NLP application and effective approach in user story research. As a new research area, it is interesting to know the direction of user story research that applies to NLP techniques. This study mainly aims to survey the state-of-the art use in NLP specifically focuses on user stories in ASD. We formulated the following research questions to fulfil these objectives:

1. RQ1: What are the uses of NLP for user stories in ASD?
2. RQ2: What are the existing NLP techniques used on user stories in ASD?
3. RQ3: What are the appropriate validation methods of NLP for user stories in ASD?
4. RQ4: What are the existing NLP tools applicable on user stories in ASD?
5. RQ5: What is the limitation of using NLP on user stories in ASD?

3.3. Search Process Strategy

We obtained relevant studies by identifying the keywords, creating a search string, and defining a database and search parameters. The set of the keyword was determined based on the objectives and research questions, specifically the uses, techniques, validation methods, tools, and limitation of using NLPs for user stories in ASD research. We identified three main categories to determine keywords based on the objectives and research questions: *'natural language processing', 'user story' and 'agile software development'*. We pinpointed alternative spelling and synonyms to acquire comprehensive results; Table 2 is listing the final set of keywords.

Table 2. Keyword used for search

Category	Keywords
Natural Language Processing	Natural Language Processing, NLP, Natural Language
User Story	User Story, User Stories
Agile Software Development	Agile Software Development, ASD, Agile Development, Agile-based Software Development

3.4. Search strings

Ensuring the formulations of accurate search strings is an essential process when performing an online search in an electronic database [26] to ensure the quality of the elicited studies. In this review, the search strings were formulated based on a stepwise procedure as low:

1. Identify the main term based on the specified research questions.
2. Finding the alternative synonyms and spelling of the identified main term.
3. Validating the search terms in any relevant research study
4. Integrating these strings with Boolean operators (AND/OR)

We make minor adjustments to the search string based on the electronic database characteristic. These adjustments were made without changing the determined set of keywords such as making this such a lower string case, applying that such item only in the research article if possible. The keywords used in search terms are applied with a combination of Boolean operators (AND/OR) to make the search process more pertinent and extend the searching process.

3.5. Resources

In this research, the relevant studies extracted from well-recognized digital libraries and scientific databases as presented in Table 4. These digital libraries were selected because they are considered relevant libraries for SLRs in SE [26]. In addition, they can provide at least one online search engine with options for conducting advanced search by keywords and results filtering via publication year and type or by domain area. The search process is compiled from various types of publication such as published conference proceedings, journal paper, chapters in books and workshops. This paper limits the publication spirit from the year 2014 to 2024. We limited the publication to ten years in hopes of obtaining the latest state-of-the-art researchers. Table 3 presented the details of the adaptations of these search string applications in the electronic database.

Table 3. Search Sources

Category	Items	
Electronic Databases	**Resource Name**	**Resource Link**
	IEEE Xplore	(http://www.ieee.org/web/publications/xplore/)
	Science Direct	(https://www.sciencedirect.com/)
	Scopus	(https://www.scopus.com/)
	Springer Link	(http://www.springerlink.com)
Document Type	Research Articles	
Source Type	Journal	
Language	English	
Access	Open Access	
Publication period	2014 to 2024	
Search applied on	Abstract, Document Tittle, Keywords	
Query Search	*"User story" OR "user stories" AND "Agile Software Development", OR "ASD" OR "Agile Development "OR "Agile-based Software Development" AND "natural language processing" OR "NLP" OR "natural language" AND PUBYEAR > 2014 AND PUBYEAR < 2024 AND (LIMIT-TO (OA , "all")) AND (LIMIT-TO (PUBSTAGE , "final")) AND (LIMIT-TO (SRCTYPE , "j") OR LIMIT-TO (SRCTYPE , "p")) AND (LIMIT-TO (LANGUAGE , "English")) AND (LIMIT-TO (DOCTYPE , "or")) AND (LIMIT-TO (SUBJAREA , "COMP"))*	

3.6. Study selection criteria

The study selection strategy is performed to determine whether the compiled studies in the initial stage of the search process must be included [25]. In this research, the study selection strategy is implemented by considering two sub-criteria: inclusion and exclusion criteria, and quality assessment criteria.

3.6.1. Inclusion And Exclusion Criteria

We used the inclusion and exclusion criteria to select relevant studies as tabulated in Table 4 below:

Table 4. Inclusion and Exclusion Criteria

No.	Study Selection Criteria	Descriptions
1.	Inclusion Criteria	(I1) Research works that are written in English-based. (I2) Research works that focus on User Stories in ASD and/ or NLP research domain from 2014 to 2024. (I3) All research works focusing on the User Stories in ASD (or, and) NLP technique based on the keywords and title of the papers. (I4) Peer-reviewed publication (I5) Relevant papers that include potential and answer to RQ by carefully examining the abstract of collected papers. (I6) All research work should be in open access mode.
2.	Exclusion Criteria	(E1) All research studies that are not written in English. (E2) Duplicates papers, excluding multiple copies of the same study and including only the most complete and recent. (E3) All papers that do not relate to the specified research questions. (E4) Paper that involves short papers, doctoral symposium papers, summary of conference keynotes, proposal, lecture notes, editorials, comments, tutorials, and review papers. (E5) Paper that is published in a predatory journal or conference. (E6) All papers considered to be grey papers which do not have bibliographic details such as publication type/date

We used the abstract title, keywords to evaluate papers based on the inclusion and exclusion criteria for initial screening. When necessary, we also opened the full text of the paper to evaluate the inclusion and exclusion criteria. We then downloaded the full text of relevant studies to assess the inclusion and exclusion criteria. We filtered out the studies that are not in compliance with the criteria. Studies that fit with our criteria were marked as primary studies. We eliminated redundant studies. With this approach, we can be more effective in choosing papers for primary studies.

3.6.2. Backward and Forward Snowballing

We used the snowballing technique to acquire more comprehensive results and reduce the risk of missing relevant studies [27]. We applied backward and forward snowballing for each identified primary study. Backward snowballing was done by examining the reference list from the primary studies to pinpoint additional papers. Forward snowballing was accomplished by examining other papers citing primary studies. Each primary study identified is a subject of further backward and forward snowballing process.

3.7. Conducting The Review

This section presents the results of the study search and selection process. We also present the quality assessment result herein.

3.7.1. Study Search and Selection

We searched the four (4) following online libraries based on the predefined search strings: IEEE Xplore, Science Direct, Scopus and Springer Link. We ran the switch on the electronic database sequentially to make this search effective. First, research Scopus recorded the results in a spreadsheet and amended. Chronologically, this search was followed by that on IEEE Xplore, Science Direct, Scopus and Springer Link. Some databases provided CSV download features that simplify this task. We ran the screening process by checking the title, abstract and keywords and applying the rules of the inclusion and exclusion criteria. Relevant papers were marked on a spreadsheet, downloaded, and included in Mendeley software. We also ensure that no redundant studies use this approach. Searches on Scopus were performed from the beginning and the last because both search engines are abstract indexing collecting data from many sources. The other databases included in the digital library category such as IEEE Xplore, Science Direct, Scopus and SpringerLink were searched between Scopus, hence the paper that appears can be easily identified in case of redundancy and reduce efforts to manage redundant papers. Papers related to research questions also have a high likelihood of being discovered in this systematic literature review. A total of 52 relevant studies were found using this method. The full text of studies was assessed for eligibility. This assessment was done by reviewing the inclusions and exclusions criteria once again and confirming whether the article was eligible for the SLR topic. The backward and forward snowballing technique would apply after discovering the primary studies. For the backwards normally we use a reference list to obtain the relevant studies. Simultaneously for the forward snowballing we checked to see the citations of these selected sites in Google Scholar. During the initial screening, we read the title of the reference or citations to decide whether the studies were relevant. We downloaded the full text of the relevant study candidates to assess them using the inclusions and exclusion criteria. 10 candidates were identified for the relevant studies. Three studies were added to the primary studies after applying the inclusions and exclusion criteria. Figure 2 presents the study search and selection process.

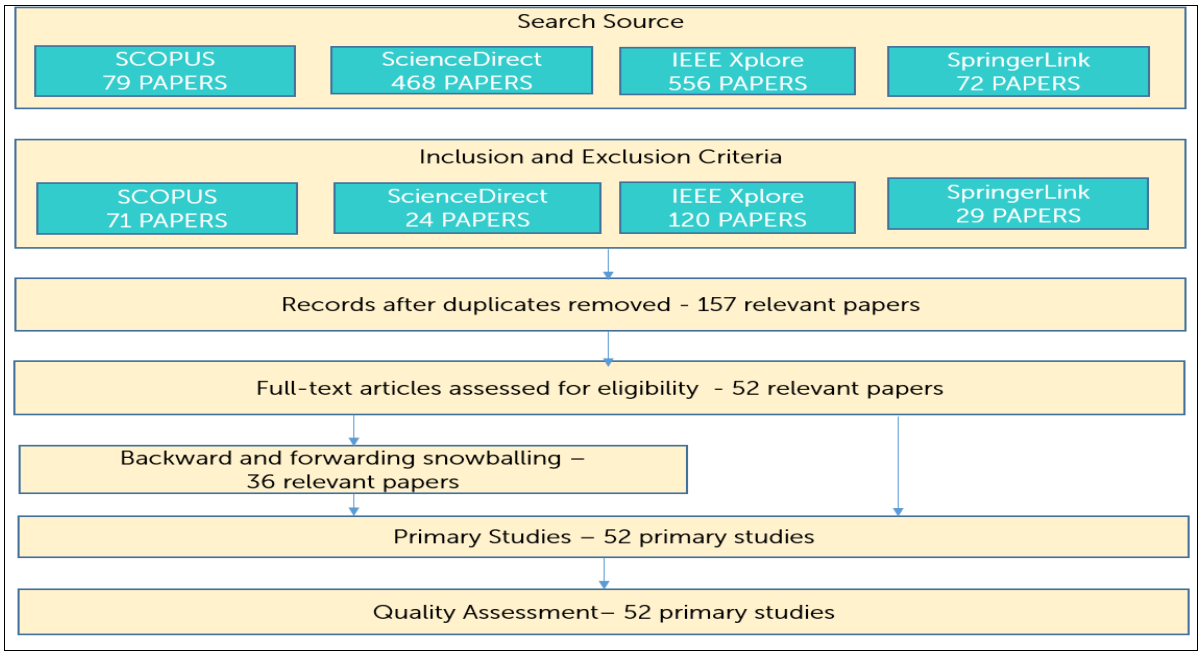


Fig 3. Study search and selection process

3.8. Quality Assessment Criteria (QAC)

We used Quality Assessment Criteria (QAC); to evaluate the methodological quality of the primary studies, we adopted the quality assessment applied by [1]. The QAC, considered one of the most essential stages in the study selection strategy [25], is executed to assess the quality of selected studies. The assessment of the selected studies is performed based on QAC quality questions [25], which were formulated based on the specified research questions that are related to our research domain. Table 6 shows the checklist to evaluate the quality of the studies included. Each QAC question has only three answers: yes, partially, and no. If a study received 'yes' as an answer, then a quality point of 1 is assigned to it, a quality point of 0.5 is assigned to a study that received 'partially' as an answer, and a quality point of 0 is awarded to a study that received 'no' as the answer. The QAC was applied with the participation of all authors of this work by precisely studying the title, abstracts, and contents of each study. First, each author assigned a quality point to each defined question. The results of the quality scores for each selected study that had been reviewed by the authors were collected. The comparisons and discussion among the authors were then conducted to address the contradictions with the purpose of obtaining a consensus in specifying the final quality score for each question and obtaining the overall quality scores of the study by summing all the quality points of the defined questions. However, to ensure the reliability of the review's findings, only the relevant studies that received a score greater than 3.5 are included, which is half of the full score (7). As a result, 52 out of 244 research studies are selected as primary studies to the research domain. The result of the quality score for each selected study is presented in Table 5. For details on the retrieval and selection process for selecting the primary studies, refer to the section on 'Document Retrieval'.

Table 5. Quality Criteria for the study selection

ID	Question(Q)	Answer Score Points	Descriptions
QA1	Was there a clear statement of the objectives of the	-1	[No]- The objective was not described
		0	[Partially] - The objective was partially but not clearly described

ID	Question(Q)	Answer Score Points	Descriptions
	research?	1	[Yes] - The objective was well described and clear.
QA2	Does the research introduce detailed descriptions of the proposed solution or approach?	-1	[No]-The details were missing
		0	[Partially] - If you wish to use the approach or solution, you may read the references.
		1	[Yes] - The approach can be used with the presented details.
QA3	Is the proposed solution or approach validated?	-1	[No] - It was not validated.
		0	[Partially] - Validated in the laboratory, or only portions of the proposal were validated.
		1	[Yes] - By case study
QA4	Does the research present an opinion or viewpoint?	-1	[No]- It is does
		0	[Partially]- Because the corresponding work was explained and the paper was set into specific paper.
		1	[Yes] - The paper is based on research.
QA5	Has the study been cited in other scientific publications?	-1	[No]- No one cited the study
		0	[Partially]- Between one and five scientific papers cited the study
		1	[Yes]- More than five scientific papers cited to the study

All primary studies, which are 52 papers were assessed on the quality assessments shown in Table 6. The first item QA1 assesses the purpose of each study. This question was answered positively in 92% of the studies. The second item QA2 is assessed if the study presents a detailed description of the approach. This question was responded to positively in 87% of the studies. The third item QA3 asks about the validation method of the result. Only 26% of the studies employed appropriate validation methods. The fourth item QA4 assessed if studies are based on the researcher rather than opinion and viewpoint. Only 28% of the studies responded positively. The final item QA5 searches for the number of citations obtained by studies, Consequently, 46% of studies were cited more than five times by other studies. Figure 3 shows the quality assessment scores for primary studies.

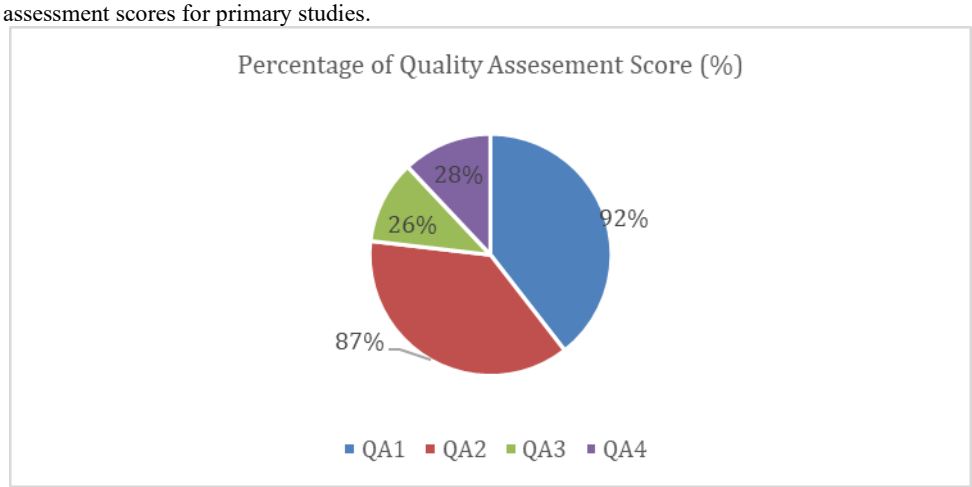


Fig 4. Percentage Score for the Quality Assessment

3.9. Data Extraction and Synthesis

The data abstraction was performed to obtain information relevant to the research question. The data was extracted following predefined extractions from Table 6. When using these tables, it enables us to record the full details of primary studies to address our research questions.

Table 6. List of attributes for analyzing primary studies

No	Study Data	Descriptions	Relevant to RQ
1.	Identifies	Unique ID for the study	Study Overview
2.	Tittle	Title of the paper	Study Overview
3.	Year	Year published	Study Overview
4.	Type of article	Journal, conferences, article, book chapter	Study Overview
5.	Application context	Industrial, academic	Study Overview
6.	Research Goal	What is the contribution of the study?	RQ1
7.	Research Goal Category	Conceptual Model extraction, software artifacts from user stories, user stories similarity, priority and size estimation, user story quality, user stories extraction	RQ1
8.	Research Methods	What research methods did the study employ?	RQ2
9.	Data	What data did the study use?	RQ2
10.	Validation	What is the validation technique doing the study apply?	RQ3
11.	NLP Technique	What NLP technique did they use for user story?	RQ2

No	Study Data	Descriptions	Relevant to RQ
12.	NLP Tools	What NLP tool that they use for user story?	RQ4
13.	Challenge and Limitation	What challenges and limitations did the study acknowledge?	RQ5
18.	Future Work	What future work did the author suggest?	RQ1

3.10. Threats to Validity

This systematic review on the Quality User Story (QUS) domain faces three primary threats: completeness, publication bias, and data synthesis. To mitigate the completeness threat, a rigorous review protocol and search strategy were employed, resulting in the selection of
52 studies were published between 2014 and 2024. However, despite these efforts, some relevant studies may have been missed, especially given the growing interest in QUS from 2016 onward, and the exclusion of non-English publications may have led to the omission of significant research. For data synthesis, the study utilized predefined Quality Assessment Criteria (QAC) to identify studies capable of adequately addressing the research questions, though there is no absolute certainty that the QAC fully achieved this goal. Publication bias, another notable threat, arises from the tendency to publish positive outcomes over negative ones. To address this, a thorough selection process and quality assessment were applied to ensure the validity and relevance of the included studies. Nonetheless, the exclusion of grey literature such as ongoing research, technical reports, and non-peer-reviewed publications represents a limitation, as it may have led to the omission of potentially valuable insights.

4. Results and Discussions

4.1. Summary of Studies

We identified 52 primary studies based on the review method. Almost half of the primary study settings were preliminary studies. 16 studies (31%) express ideas and present it at the very least, experimentations or a case study as proof of concept 36 studies (69%) use an inlet academic setting for research. No studies used industry settings, however several use real data sets from the industry in their research.

4.2. RQ1 What are the uses of NLP for User Stories?

The results of the primary study illustrated several natural language applications in user stories. We use the category of NLP RE tools [28] to classify the goal of the primary studies: discovering the defects, generating a model or artifact, tracing the link between model and natural language requirements and identifying the key extractions. Table 8 presents a summary of the primary studies based on this category. Figure 4 below illustrates the year-wise distributions of the categorized primary study goals. Two topics are the major concerns that took most of the researcher's attention: identifying the key abstractions and generating models/artifacts. Both topics have continued to be studied on an ongoing basis since 2014. The topic of abstraction identification became the primary choice in the early phases because researchers are still trying to gain an understanding of the new and different characteristics of user stories. The topic of generating models' artifacts is always a challenge in software engineering research because it can accelerate the software development time. The following subsections present the directions of research and that by primary studies for each category.

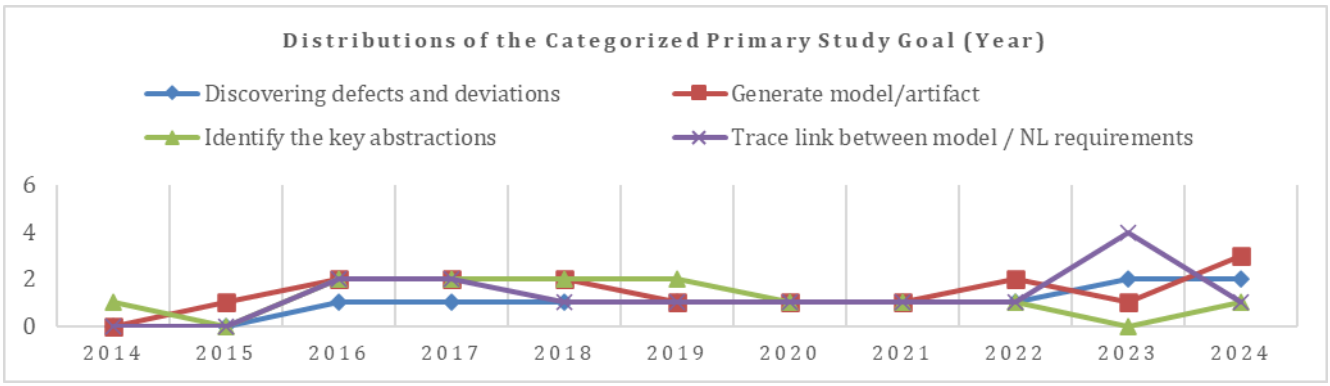


Fig 5. Distributions of the categorized primary study goal by year

4.3. RQ2 What approaches were available in research related to NLP in User Stories?

To answer research question two about the approaches available in research related to the NLP in user stories divided into several pieces which are NLP techniques, validation methods and tools used. Table 7 below shows the NLP technique in user story studies in detail.

Table 7. NLP Technique in user story studies

NO	NLP METHODS	PURPOSE
1.	Preprocessing	It is for treating data into the desired form; the process usually includes tokenization filtering and stop word removal.
2.	Part of Speech (POS) Tag	Identify the lexical categories in a sentence, such as nouns, adjectives, and adverbs. POS accurately detects noun and verb phrases, which helps researchers find key parts of a user story: "who" (usually noun phrases), "what" and "why" (typically verbs followed by noun phrases).
3.	Named entity	It's a technique of word finding and classifying named entities in unstructured text. They were usually

NO	NLP METHODS	PURPOSE
	recognizer (NER)	used to identify people, organizations or other entities written in the text. Semantic role labeling is the process of assigning a label to a word or phrase in sentences indicating its semantic role.
4.	Bag of words	It is a technique of grouping words and calculating the other term frequency to measure their level of importance
5.	Machine Learning	To automate, enhance, and analyze the management of user stories in software development. This can improve efficiency, quality, and decision-making during the Agile process.
6.	Clustering	Process of automatically grouping similar user stories together based on their content. This helps product teams organize, prioritize, and manage large backlogs more efficiently
7.	Term frequency – inverse document frequency	Identify important keywords in each user story and distinguish unique terms that differentiate one story from others which help in searching/filtering user stories in large backlogs, clustering/grouping similar user stories, prioritizing or tagging stories by relevant topics and feature extraction for machine learning models
8.	Lemmatization	Is the process of grouping word form to be analyzed in one item dictionary form another similar approach is stemming we change to each raw form.
9.	Semantic role labeling	Identifies who did what to whom, when, where, and how in a sentence. It breaks down the predicate-argument structure of a sentence to assign semantic roles (like Agent, Patient, Instrument, etc.) to words or phrases.
10.	Similarity matrix	Measure how semantically similar different user stories are to each other. This helps in identifying duplicates or near-duplicates, related stories that may belong to the same epic, redundant features and grouping stories into themes or components
11.	Dependency	Is the activity of extracting dependencies from and sanctions that represent grammatical structure and defining the relationships between words.
12.	Open information extraction	Automatically extract structured information (typically triples) from unstructured text, without requiring a predefined schema or ontology.
13.	Syntactic parse tree	Visual or structured representation of the grammatical structure of a sentence based on formal grammar (usually constituency grammar or dependency grammar).

4.4. RQ3 Validation Methods

We examined 4 types of validation conducted by researchers to assess the result, precision and recall, case study/example, average time and effort comparison and prototype demonstration. Many primary studies employ case studies for evaluation methods. This evaluation method reports experience based on the best examples which usually provide lessons learned. Besides, several studies used prototype demonstration as proof of their concept. Several other studies conducted evaluations by comparing the tools' performance with control elements such as the average time and effort required by tools compared to the group of experts. The evaluations of studies in the NLP field usually employed precision, recall and F-measure as the quality indicators. Precision is how many of the items selected are relevant, as shown in (1). A recall is how many relevant items are selected as shown in (2). F-measure is the unit precision and recall as shown in (3). Unexpectedly, the evaluations using precision and recall are not main evaluations conducted by the primary studies. Figure 4 below shows the user stories validation methods calculation.

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (1)$$

$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (2)$$

$$\text{F-measure} = 2\ x\ \frac{Precision * Recall}{Precision + Recall} \quad (3)$$

Notes:
True Positive = The correctly labeled instances
False Negative = Incorrectly labeled instances
False Negative = The missed-out instances by the system

Fig 6. User stories validation methods calculation

There are 4 methods of validation user stories which are case study, precision and recall, average time and effort comparison and prototype demonstration. The evaluation was done by comparing the results with the predictions made by human annotators and usually using a group of software developers or university students. What was evaluated depended on the study purpose. Most of the datasets used by researchers were independently collected and privately stored for internal needs.

4.5. RQ4 NLP Tools

Most studies used SpaCy or Standford CoreNLP to conduct the NLP. Some started using word2vec, WordNet, LinePipe Toolkit, PropBank, TreeTagger and Standford POS Tagger while some did not report what tools they utilized. More than one tool was used in some studies such as SpaCy and NLTK. Table 8 shows that list of the NLP toolkits used in the studies. The feature in the widely used tool is the POS tag which is available in almost all tools. This feature is very useful in user story study because it can be used to chunk phrases into verbs and nouns to quickly determine the aspects of who, what and why in the user's story. Also, most tools support preprocessing natural language as basic functionality, making it easier for researchers to carry out their research. Another useful feature is to calculate similarity. Word2vec is the most widely used similarity calculation implementation, besides SpaCy and WordNet also provide similar functionality with different implementation techniques.

Table 8. NLP tool in the user studies.

No	Tools	Features
1.	SpaCy	Tokenization, Part of Speech (POS) Tagging, Dependency Parsing, Lemmatization, Similarity
2.	Standford CoreNLP	Tokenization, Part of Speech (POS) Tagging
3.	Natural Language ToolKit (NLTK)	Part of Speech (POS) Tagging

No	Tools	Features
4.	word2vec	Semantic Arithmetic
5.	WordNet	Semantic Similarity
6.	LingPipe Toolkit	end-of-sentences detection
7.	PropBlank	Semantic Propositions
8.	TreeTagger	Part of Speech (POS) Tagging
9.	Standofrd POS Tagger	Part of Speech (POS) Tagging

4.6. RQ5: What are the challenges or limitations of using NLP in user story research?

The primary studies reported several challenges. Some were related to the improvement of recall and precision, dataset, understanding the correct interpretation of a sentence, and human intervention. Table 9 is the summarization of challenges reported in the primary studies.

Table 9. Challenges

Challenges	Descriptions	Findings
Improving the recall and precision	Low precision	Obtained low precision values and Obtained consistent recall results above 90% but the average precision value was still approximately 72-77%
Dataset	Heterogeneity and low amount of data and manual data tagging	Detecting the ambiguity of user stories can be time consuming even though it has been done using tools
Context / domain dependent	Cannot be generally used in all contexts of the problem	Ambiguity in Natural Language such Words or phrases have different meanings across domains.
	Not yet able to handle complex systems	Ontology and Terminology Variation such each domain has its own vocabulary and relationships.
Understanding the correct interpretation of a sentences	Compounds are difficult to identify correctly; conjunctions are also a challenge and same verb but can be classified in different categories	Understanding the proper sentence interpretation remains a challenge
Human Intervention	Complement rather than replace human decision making	Results obtained cannot yet match with human results. The NLP implementation on the software requirements usually cannot fully implement automation but this can be accomplished in software development.

5. Conclusion

We observed that Europe is still the center of research in this area. Many primary studies from Europe have become references to other primary studies. The geographic location distribution is a good signal for the research area development. Studies on the NLP and user stories are already the content of researchers from different countries. More than half of his studies were preliminary studies indicating that the research area is not mature and still at its early stage. This is normal because ASD is a research field and it's also newly developed. The number of publications in this area increases every year. The conference and book chapters still dominate the publication area. This is natural for a new and emerging field of science because the conference and the book chapter offer a relatively fast process in a publication compared to journals. The year 2016 has also begun publication in journals that marked the improvement in research quality.

The primary focus of NLP in user story research has been on identifying abstractions and generating models, as researchers are still exploring the structure and characteristics of user stories, which are semi-structured and relatively easy to analyze. Efforts have been made to define ontologies and semantic relationships among user stories for goal-based grouping, but little work has addressed how to extract user stories from unstructured free text. This task remains challenging due to the complexity of natural language and the need to analyze language structure deeply to identify key aspects like who, what, and why. When free text is sourced from software-related documents—such as app reviews, user comments, or description extraction techniques can be adopted, while data from non-software sources like news or social media requires additional filtering to determine relevance. Named Entity Recognition (NER) helps extract the "who," and identifying causal relationships can clarify the "why." Researchers commonly extract nouns and verbs from user stories to generate software artifacts such as class diagrams, sequence diagrams, use case diagrams, and BPMN, often using model patterns and predefined rules. Recently, research has shifted toward detecting defects and tracing links between model artifacts, supported by techniques like machine learning and semantic similarity. Although user stories serve multiple functions, documentation, artifact generation, and validation—challenges remain in ensuring their quality, as natural language-based requirements often suffer from issues of consistency, completeness, and correctness, which some studies have begun to address.

Most current research on using Natural Language Processing (NLP) in user stories focuses on basic techniques like preprocessing and part-of-speech (POS) tagging to extract verbs and nouns, which support identifying the "who, what, and why" in requirements. However, the semantic dimension of NLP remains underexplored, presenting an opportunity to enhance its role in user story research. Approaches like deep learning and semantic analysis (e.g., using bags-of-concepts or narratives) are promising but limited by the lack of large, open-access datasets, as user stories are often proprietary and not shared due to privacy concerns. This restricts progress and calls for the development of high-quality open datasets. User stories vary in scope and structure, especially epics with sub-stories, making clear definition and traceability crucial when generating software artifacts. Despite the growing use of advanced techniques like machine learning, clustering, semantic role labeling, and similarity calculations, most studies remain preliminary, lack detailed methodological descriptions, and often rely on case studies instead of rigorous evaluations using metrics like precision and recall.

References

[1] E. M. Schön, J. Thomaschewski, and M. J. Escalona, "Agile Requirements Engineering: A systematic literature review," *Comput. Stand. Interfaces*, vol. 49, pp. 79–91, 2017, doi: 10.1016/j.csi.2016.08.011.

[2] R. Noel *et al.*, "Exploring collaborative writing of user stories with multimodal learning analytics: A case study on a software engineering course," *IEEE Access*, vol. 6, pp. 67783–67798, 2018, doi: 10.1109/ACCESS.2018.2876801.

[3] Y. Wautelet, S. Heng, M. Kolp, and I. Mirbel, *Unifying and extending user story models*, vol. 8484 LNCS. 2014.

[4] G. Lucassen, F. Dalpiaz, J. M. E. M. van der Werf, and S.Brinkkemper, "Improving agile requirements: the Quality User Story framework and tool," *Requir. Eng.*, vol. 21, no. 3, pp. 383–403, 2016, doi: 10.1007/s00766-016-0250-x.

[5] I. Inayat, S. Salwah, S. Marczak, M. Daneva, and S.Shamshirband, "A systematic literature review on agile requirements engineering practices and challenges," *Comput.Human Behav.*, vol. 51, pp. 915–929, 2015, doi: 10.1016/j.chb.2014.10.046.

[6] M. Younas *et al.*, "Elicitation of Nonfunctional Requirements in Agile Development using Cloud Computing Environment," *IEEE Access*, pp. 1–1, 2020, doi: 10.1109/access.2020.3014381.

[7] Pasaribu, J. S., & Argadikusuma, I. S. (2024). *Design and testing of a web-based student information management system*. *International Journal of Engineering, Science and Information Technology*, *4*(4). https://ijesty.org/index.php/ijesty/article/view/594

[8] Karyono, K., Violin, V., Osman, I., Rao, D. G., & Apramilda, R. (2024). *Analysis of the interrelationship of human resource performance, digital service quality, perceived of service value and customer loyalty*. *International Journal of Engineering, Science and Information Technology*, *4*(3). https://ijesty.org/index.php/ijesty/article/view/527

[9] Yunus, A., Wilanda, A., Haji, W. H., Alatas, A. R., & Dharmawan, D. (2024). *Analysis of the influence of training and capacity development programs on improving service quality and performance of medical personnel in handling patients*. *International Journal of Engineering, Science and Information Technology*, *4*(2), 11–15. https://ijesty.org/index.php/ijesty/article/view/484

[10] H. Meth, M. Brhel, and A. Maedche, "The state of the art in automated requirements elicitation," *Inf. Softw. Technol.*, vol. 55, no. 10, pp. 1695–1709, 2013, doi: 10.1016/j.infsof.2013.03.008.

[11] A. R. Da Silva, "Linguistic patterns and linguistic styles for requirements specification (I): An application case with the rigorous rsl/business-level language," *ACM Int. Conf. Proceeding Ser.*, vol. Part F1320, no. I, pp. 1–27, 2017, doi: 10.1145/3147704.3147728.

[12] H. Dar, M. I. Lali, H. Ashraf, M. Ramzan, T. Amjad, and B. Shahzad, "A systematic study on software requirements elicitation techniques and its challenges in mobile application development," *IEEE Access*, vol. 6, pp. 63859–63867, 2018, doi: 10.1109/ACCESS.2018.2874981.

[13] T. Johann, C. Stanik, A. M. B. Alizadeh, and W. Maalej, "SAFE: A Simple Approach for Feature Extraction from App Descriptions and App Reviews," in *Proceedings - 2017 IEEE 25th International Requirements Engineering Conference, RE 2017*, 2017, pp. 21–30, doi: 10.1109/RE.2017.71.

[14] I. K. Raharjana, F. Harris, and A. Justitia, "Tool for Generating Behavior-Driven Development Test-Cases," *J. Inf. Syst. Eng. Bus.Intell.*, vol. 6, no. 1, p. 27, Apr. 2020, doi: 10.20473/jisebi.6.1.27-36.

[15] Y. Wautelet, S. Heng, S. Kiv, and M. Kolp, "User-story driven development of multi-agent systems: A process fragment for agile methods," *Comput. Lang. Syst. Struct.*, vol. 50, pp. 159–176, 2017, doi: 10.1016/j.cl.2017.06.007.

[16] W. Behutiye *et al.*, "Management of quality requirements in agile and rapid software development : A systematic mapping studying. *Softw. Technol.*, vol. 123, no. April 2019, p. 106225, 2020, doi: 10.1016/j.infsof.2019.106225.

[17] S. Tiwari and A. Gupta, "A systematic literature review of use case specifications research," *Inf. Softw. Technol.*, vol. 67, pp.128–158, 2015, doi: 10.1016/j.infsof.2015.06.004.

[18] G. Loniewski, E. Insfran, and S. Abrahão, "A systematic review of the use of requirements engineering techniques in model-driven development," in *Model Driven Engineering Languages and Systems. MODELS 2010. Lecture Notes in Computer Science*, vol. 6395,2017, LNCS, pp. 213–227, doi: 10.1007/978-3-642-16129-2_16.

[19] P. Heck and A. Zaidman, *A systematic literature review on quality criteria for agile requirements specifications*, vol. 26, no. 1. Springer US, 2018.

[20] 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, doi:10.1016/j.jss.2015.05.006.

[21] F. Nazir, W. H. Butt, M. W. Anwar, and M. A. Khan Khattak,"The Applications of Natural Language Processing (NLP) for Software Requirement Engineering - A Systematic Literature Review," in *Information Science and Applications 2017. ICISA 2017. Lecture Notes in Electrical Engineering*, 2017, vol. 424, no.March 2017, pp. 485–493, doi: 10.1007/978-981-10-4154-9_56.

[22] B. Kitchenham and S. Charters, "Guidelines for performing Systematic Literature Reviews in Software Engineering," 2007.

[23] A. Pollock and E. Berge, "How to do a systematic review," *Int. J.Stroke*, vol. 13, no. 2, pp. 138–156, 2018, doi:10.1177/1747493017743796.

[24] A. Liberati *et al.*, "The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration.," *BMJ*, vol.339, 2009, doi: 10.1136/bmj.b2700.

[25] H. Villamizar, M. Kalinowski, A. Garcia, and D. Mendez, "An efficient approach for reviewing security-related aspects in agile requirements specifications of web applications," *Requir. Eng.*, vol. 25, no. 4, pp. 439–468, 2020, doi: 10.1007/s00766-020-00338-w.

[26] R. Barbosa, A. E. A. Silva, and R. Moraes, "Use of Similarity Measure to Suggest the Existence of Duplicate User Stories in the Srum Process," *Proc. - 46th Annu. IEEE/IFIP Int. Conf. Dependable Syst. Networks, DSN-W 2016*, pp. 2–5, 2016, doi: 10.1109/DSN-W.2016.27.

[27] D. Badampudi, C. Wohlin, and K. Petersen, "Experiences from using snowballing and database searches in systematic literature studies," *ACM Int. Conf. Proceeding Ser.*, vol. 27-29-Apri, no.April, 2015, doi: 10.1145/2745802.2745818.

[28] D. Berry, R. Gacitua, P. Sawyer, and S. F. Tjong, "The case for dumb requirements engineering tools," *Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 7195 LNCS, pp. 211–217, 2012, doi: 10.1007/978-3-642-28714-5_18.

[29] F. Dalpiaz, A. Ferrari, X. Franch, and C. Palomares, "Natural Language Processing for Requirements Engineering: The Best Is Yet to Come," *IEEE Softw.*, vol. 35, no. 5, pp. 115–119, 2018, doi: 10.1109/MS.2018.3571242.

[30] K. Athiththan, S. Rovinsan, S. Sathveegan, N. Gunasekaran, K. S. A. W. Gunawardena, and D. Kasthurirathna, "An Ontology-based Approach to Automate the Software Development Process," 2018, doi: 10.1109/ICIAFS.2018.8913339.

[31] F. S. Bäumer and M. Geierhos, "Running out of words: Howsimilar user stories can help to elaborate individual natural language requirement descriptions," in *Communications inComputer and Information Science*, 2016, vol. 639, pp. 549–558,doi: 10.1007/978-3-319-46254-7_44.