Journal of
Information Systems Engineering
and Business Intelligence

Vol.10, No.2, June 2024
Available online at: http://e-journal.unair.ac.id/index.php/JISEBI

Text Stemming and Lemmatization of Regional Languages in Indonesia: A Systematic Literature Review

Zaenal Abidin [1], **Akmal Junaidi** [2]*, **Wamiliana** [3]

[1]*Doctoral Program of Mathematics and Natural Sciences, Universitas Lampung, Bandar Lampung, Indonesia*
[1]2237061006@students.unila.ac.id

[1]*Faculty of Engineering and Computer Science, Universitas Teknokrat Indonesia, Bandar Lampung, Indonesia*
[1]zabin@teknokrat.ac.id

[2)3)]*Faculty of Mathematics and Natural Sciences, Universitas Lampung, Bandar Lampung, Indonesia*
[2]akmal.junaidi@fmipa.unila.ac.id, [3]wamiliana.1963@fmipa.unila.ac.id

Abstract

Background: Stemming is significantly essential in natural language processing (NLP) due to the ability to minimize word variations to fundamental forms. This procedure facilitates the analysis of textual data and enhances the precision of classification and information retrieval.
Objective: Previous related systematic literature review has not been conducted on stemming and lemmatization in regional languages in Indonesia. Therefore, this study aims to conduct a systematic literature review to capture the latest developments in stemming and lemmatization in regional languages in Indonesia.
Methods: This study was carried out using Kitchenham method, analyzing 35 studies extracted from 740, which were obtained from Scopus, IEEE Xplore, and Google Scholar, and published between 2014 and 2023.
Results: The results showed that study trends in stemming possessed the potential to continue developing every year. Additionally, the main element in stemming and lemmatization studies was found to be the availability of digital dictionaries in regional languages. This was because greater number of basic vocabularies contributed more positively to stemming or lemmatization. The availability of word morphology information in regional languages would be constructive for making rule-based stemmers. Meanwhile, corpus-based stemming and lemmatization studies could only be conducted for languages with a large corpus to ensure there were various affixed words to process.
Conclusion: Based on SLR study, stemming and lemmatization in regional languages in Indonesia developed significantly from 2014 to 2023. The two main strategies applied included using available digital dictionaries and language morphology information. However, the main challenges encountered were the limited number of vocabulary words in the dictionaries and testing various rule-based methods.

Keywords: Lemmatization, Morphology, Rule-based, Stemming, Systematic Literature Review.

Article history: Received 4 October 2023, first decision 17 March 2024, accepted 3 June 2024, available online 28 June 2024

I. INTRODUCTION

The natural language processing to text stemming is a method of minimizing words to their roots [1] [2] [3]. In addition to the general application, this method works well in languages with complex word-formation systems, such as Indonesian [4]. Text stemming is essential for many text analysis and information retrieval applications in natural language processing (NLP) [5]. It eliminates prefixes and suffixes to obtain the root, or "stem", by simplifying sentences to their basic meaning, as well as facilitating text processing, search, and analysis [1] [2] [3].

The relevance of text stemming is the ability to enhance information retrieval and natural language processing tasks. By reducing words to their root forms, stemming facilitates the consolidation of related words, improving search accuracy, and reducing redundancy [5]. This process aids in overcoming challenges posed by variations in word forms, such as plurals, verb tenses, and derivations. Consequently, text stemming is crucial in several applications, including search engines, document clustering, sentiment analysis, and machine translation [1][2][5]. The implementation also

[*] Corresponding author

doi: http://dx.doi.org/10.20473/jisebi.10.2.217-231

enables more efficient and practical analysis of text normalization methods, offering several benefits in the academic domain.

Some of these benefits include fewer differences in text, better retrieval, more accurate text classification, reduced storage space usage, easier sentiment analysis, normalized text, decreased noise in text data, improved information extraction processes, and simplified language processing tasks [1] [2] [5]. Stemming and lemmatizing are often related, as these strategies simultaneously function to reduce input text diversity [1].

Language preprocessing methods such as stemming and lemmatization prevent the omission of several forms of words [6]. Preprocessing transforms the data into a format that is easily understood, consistent, and analyzable. Information Retrieval (IR) preparation uses stemming and lemmatization as crucial procedures, categorizing words that are derived from the same root or have the identical canonical citation form under a common fundamental idea [1] [6].

Various studies have been conducted on stemming in Indonesia since 1999, starting from Nazief and Adriani method to the UG 18 algorithm in 2019 [7]. These studies have produced more than 40 publications on Indonesian stemming-based studies [7] [8], while investigation on regional languages started to appear approximately in the last 10 years. Indonesia possesses a rich culture with numerous rituals, customs, performances, and linguistic diversity, comprising over 700 languages spoken across the archipelago [9]. The country also has the second-largest multilingual population, speaking approximately 700 languages, including Javanese, Sundanese, Madurese, Balinese, Minangkabau, and Riau Malay.

Bahasa Indonesia, a standardized form of Malay, serves as the national language and the main means of communication [10] [11]. The Indonesian language borrows lexically from Javanese, Sundanese, Minangkabau, Dutch, Sanskrit, Portuguese, Arabic, and English [9][12][13]. The country has a variety of regional languages in distinct locations, [9], which are used by locals for daily communication and culture.

Some Indonesian languages have separate lexicons, syntax, and cultural importance, with regional languages playing a crucial role in the cultural diversity and history. Regional languages help preserve and promote the unique cultural identities of many areas and ethnic communities [9], conveying local literature, folklore, rituals, and interpersonal contact. Despite schools teaching Bahasa Indonesia and its use in formal settings, regional languages are essential for preserving the rich linguistic and cultural history [9][12][13]. Therefore, the success of stemming studies in Indonesia [6] [7] [8] has inspired the investigation of regional languages.

Based on the description in [6] [7] [8] , the study aimed to conduct a systematic literature review on Indonesian language both in formal and non-formal languages. Although previous studies have focused on the scientometric perspective [7], there is no review of regional languages stemming from Indonesia. This shows the need to conduct systematic literature review (SLR) studies on regional languages stemming from Indonesia to discover the development of regional languages stemming from studies in Indonesia as SLR can provide an overview of the related studies conducted, including the methods used, results, and limitations, identify gaps in regional languages stemming from Indonesia as SLR can facilitate the identification of areas that have not been widely investigated or require improvement, and provide recommendations for regional language stemming studies from Indonesia to address the identified knowledge gaps.

This study aimed to identify possibilities and challenges for stemming studies on additional regional languages. The investigation was carried out using questions about stemming and lemmatization to evaluate literature and discover future opportunities. Moreover, relevant studies on SLR were obtained using Scopus, IEEE Xplore, and Google Scholar.

II. LITERATURE REVIEW

A. Stemming and Lemmatization

Stemming and lemmatization are essential preprocessing steps in natural language processing that transform words into roots. Specifically, stemming words to their roots preserves semantic meaning and removes suffixes. For example, it shortens words such as play, played, playful, player, playground, and playlist to "play" to identify their root, ignoring syntax. Lemmatization forms valid words [1] [2] [3] by reducing variability through the elimination of inflectional endings and creating dictionary words. To reach the lemma, suffixes are removed or replaced, such as "running," "run," and "ran". Word stemming eliminates prefixes and suffixes, while in a textual context, lemmatization improves meaning [1] [2] [3]. Stemming shortens 'studies' ('-es') to 'studi', but lemmatization seeks to determine the singular or third-person meaning of 'studies'. Since "studies" is the present tense of "study," it becomes the derived lemma.

B. Methods for Evaluating Stemming and Lemmatization

The three main stemming and lemmatization evaluation methods are direct, gold standard, and indirect [1] [2] [3]. Direct assessments examine stemmer ability on words without considering application based on error rates, correctly stemmed words, statistical methodologies, and other criteria. Obtaining language-specific test words requires a significant amount of physical labor. The method of Paice and Sirsat includes direct assessment, where professionals personally verify the system's accuracy by manually checking the output against the provided input. To calculate stemmer accuracy, Equation (1) uses the gold standard, which is only effective for small datasets. This method ignores already-stemmed words and shows the ratio of right stems produced by stemmer. Before information retrieval and text classification, indirect methods use stemmer to test their performance. The main benefit of these methods is the use of automated technology to check performance without human intervention, requiring query sets and document collections. Indirect methods evaluate stemmer's information retrieval using precision, F-score, recall, average precision, and mean average precision.

$$accuracy = \frac{total\ number\ of\ correct\ stem\ obtained}{total\ number\ of\ words\ given\ to\ stemmer}\ x\ 100 \qquad (1)$$

The initial step in evaluating stemmer's performance is to identify numerous faults [1] [2] [3] such as the timing, causes, and impacts of errors. Under Stemming Errors (USE) occur when stemmer removes letters below limits, resulting in words without a stem, or where affix removal alters the intended meaning. When stemmer removes several characters, over-stemming errors (OSE) occur leading to invalid stems or out-of-vocabulary OOV terms. Mis-stemming errors (MSE) occur when stripped characters fail to attach correctly, thereby distorting the word's original meaning.

C. Related Secondary Studies

SLR on stemming and lemmatization produced three studies, where the first focuses on information retrieval, specifically sentence similarity that does not use Indonesian text [6]. The second SLR study focuses on the scientometric analysis of Indonesian stems [7], but the third presents a non-formal analysis of Indonesian stemming [8]. However, Indonesia has not produced SLR in regional languages stemming.

This current internet-dependent society relies on Information Retrieval (IR) to quickly obtain relevant information [6]. Sentence similarity is important in IR, as the performance improves tasks with stemming and lemmatization. Due to the need for optimal strategy, this study aims to investigate the best preprocessing method through SLR on stemming and lemmatization drawn from previous investigations. Various evaluation methods have shown that many factors determine the selection between stemming or lemmatization as best preprocessing method. Based on the results, lemmatization outperforms stemming for sentence similarity tasks, according to the authors. Meanwhile, stemming is the best for speed optimization due to the superior computing performance.

TABLE 1
COMPARISON OF THREE STUDIES SLR

Studies	Objectives	Methods	Key Findings	Conclusions
Systematic Literature Review of Stemming and Lemmatization Performance for Sentence Similarity	Investigate the impact of stemming and lemmatization on sentence similarity tasks	Systematic literature review of studies evaluating stemming/lemmatization for sentence similarity	Stemming generally improved sentence similarity more than lemmatization, but results varied across datasets and algorithms.	Both methods can enhance sentence similarity, but stemming may be more effective. Further study is needed on advanced methods.
Stemming Algorithm for Indonesian Language: A Scientometric View	Analyze publication and citation trends related to Indonesian language stemming algorithms	Scientometric analysis using Scopus database to collect and analyze stemming algorithm publications	Indonesian stemming study grew significantly from 1999-2019, with Nazief and Adriani and Enhanced Confix Stripping (ECS) algorithms most widely studied.	Although Indonesian stemming study is advancing, there are still opportunities for new regional language algorithms and evaluations.
A Systematic Literature Review of Stemming in Non-Formal Indonesian Language	Review existing methods for stemming non-formal/colloquial Indonesian text	Systematic literature review focused on stemming methods for informal Indonesian used in social media, messaging, etc.	Most works used rule-based or hybrid methods combining rules and statistics. Few optimal solutions are required for highly irregular informal text.	Non-formal Indonesian stemming is an emerging area lacking robust, widely accepted solutions. More studies are needed on data-driven and contextual methods.

Stemming eliminates prefixes and suffixes to find a word's root, serving as a process applicable in all languages, including Indonesian. Originating from Malay and Sanskrit, Indonesian has observed the development of root-word dictionary-based stemming method, known as Nazief and Adriani method or Confix Stripping (CS) and Enhanced CS. Scientometrics tracks academic citations and scientific topics [7], focusing on the library and other important applications. This study examined studies using stemming algorithms and was selected through keywords such as Nazief and Adriani, including ECS. A powerful scientific database, Dimension.ai, delivered the documents, and bibliometric analysis was conducted using co-authorship and citation methodologies in VOS Viewer. The referenced documents were analyzed, reviewing 310 "Nazief and Adriani" and 119 "ECS" publications aiding future studies, which are Indonesian most cited and co-authored.

This study critically reviews stemming studies in non-standard Indonesian [8] to understand how data are obtained, evaluated, and interpreted. SLR was used to discover, filter, and assess studies obtained from ScienceDirect, IEEE, arXiv, the ACM Digital Library, Semantic Scholar, Google Scholar, Springer, and Elsevier, published between 2016–2022. The area of interest in Indonesian stemming included data gathering, stemming methodologies, and historical studies. This study selected 27 Indonesian stemming studies from a total of 47 published studies and the data collection process was examined, including languages stemming methods, and results.

The three SLR confirm the absence of SLR on stemming and lemmatization in regional languages in Indonesia as a measurable gap. Table 1 shows a comparison of three published studies reviewing stemming and lemmatization.

III. Methods

In this study, two stages were included in conducting SLR, where the first was planning and conducting a literature review. Meanwhile, the second is reporting the literature review results, with details of each process explained in sub-sections A to B. This SLR primarily concentrates on the regional languages prevalent in Indonesia for communication, both orally and in writing, on different islands in the country. These languages are often used by residents of the islands of Papua, Kalimantan, Sumatra, Sulawesi, Java, Bali, Madura, Timor, Halmahera, Seram, Sumbawa, Flores, and others. Moreover, regional languages are a fascinating topic in the field of natural language processing (NLP). With over 700 regional languages in the country, there is great potential for further exploration [9] [14]. SLR was conducted using a method similar to the previous study [15][16] and Kitchenham method [17][18].

A. Plan and Conduct a Literature Review

The literature review planning stage consists of determining objectives and study questions. This is followed by determination of the search strategy, inclusion and exclusion criteria to ensure that all similar studies are included in the literature review list. Studies related to stemming are increasingly widespread, with each language having different characteristics. Therefore, SLR is needed to address stemming and lemmatization specifically for regional languages, as the method must be adapted. A total of three questions are formed to fulfill SLR objectives, which include Research Question 1 (RQ1) to Research Question 3 (RQ3). The PICOC formula [15][16] was used to establish the questions in this study, as shown in Table 2. SLR does not conduct comparison, thereby the comparison category (C) value is designated as not applicable (n/a).

TABLE 2
CRITERIA OF RESEARCH QUESTION

Criteria	Elements of Searchable Question
Population (P)	Task Text Pre-processing in Natural Language Processing
Intervention (I)	Text Stemming, Text Stemmer, Text Lemmatization, Text Lemmatizer
Comparison (C)	n/a
Outcomes (O)	Stemming or lemmatization method's accuracy value and applicability in other NLP tasks.
Context (C)	Text Stemming, Text Stemmer, Text Lemmatization, Text Lemmatizer in Indonesian Regional Languages

By narrowing the scope, the questions that ensure examination in this study can be identified:

RQ1: What are the advantages and disadvantages of published studies on stemming and lemmatization of Indonesian regional languages?

RQ2: What methods do text stemming and lemmatization studies apply in Indonesia for texts originating from regional languages?

RQ3: What are the challenges of text stemming and lemmatization in regional languages in Indonesia?

Scopus, IEEE Xplore, and Google Scholar were used to conduct searches for studies on question creation. Search strategies were formed by carefully selecting the essential phrases from the topic and incorporating alternative terms and synonyms. The search query used was ("nazief adriani" AND “stemming”) OR ("enhanced confix stripping" AND “stemming”) OR ("rule-based" AND “stemming”) OR (“morphology-based” AND “stemming”) OR (“local language” AND “stemming”) OR ("nazief adriani" AND “stemmer”) OR ("enhanced confix stripping" AND “stemmer”) OR ("rule-based" AND “stemmer”) OR (“morphology-based” AND “stemmer”) OR (“local language” AND “stemmer”) OR ("nazief adriani" AND “lemmatization”) OR ("enhanced confix stripping" AND “lemmatization”) OR ("rule-based" AND “lemmatization”) OR (“morphology-based” AND “lemmatization”) OR (“local language” AND “lemmatization”) OR ("nazief adriani" AND “lemmatizer”) OR ("enhanced confix stripping" AND “lemmatizer”) OR ("rule-based" AND “lemmatizer”) OR (“morphology-based” AND “lemmatizer”) OR (“local language” AND “lemmatizer”). The query was designed based on reference [6] [7] [8], while studies on stemming in Indonesia were inspired by Nazief and Adriani and Enhanced Confix Stripping (ECS) methods that focused on rule-based, morphology-based, and regional languages.

The use of the inclusion and exclusion criteria in Table 2 aimed to maintain the concentration of the search results on the specific area of interest. The inclusion and exclusion criteria for this study were formulated based on the objectives to ensure the reliability of the gathered data. Table 3 presents the two primary criteria used in the selection method, namely inclusion and exclusion.

In the primary section of the review, the information obtained was used for relevant search comprehensively. Subsequently, selection and filtering criteria were used to identify studies included in the analysis. This section presents the results of the search and selection process, including the conclusions of the quality evaluation. Scopus, IEEE Xplore, and Google Scholar were used as the primary databases for the search, which was conducted concurrently to expedite the process, based on the predetermined search query and criteria. Each title and abstract in the search result was evaluated and correlated with the study criteria. After showing a positive correlation, the studies were added to the library for additional filtering, as shown in Fig. 1.

TABLE 3
CRITERIA OF SELECTION STUDIES CRITERIA

Criteria	Inclusion Criteria	Exclusion Criteria
Language	Written in English or Indonesian	Written in a language other than English or Indonesian
Date	2014 – 2023	Before 2014
Type of Publication	Journal articles, Conference Proceeding, Thesis or Dissertation	In addition to journal articles, conference proceedings, thesis or dissertation
Type of Studies	Stemming or lemmatization studies on regional languages in Indonesia were thoroughly examined and explored. Full-text studies.	Studies not analyzed or thoroughly studied are stemming or lemmatization studies in languages other than regional languages. Not Full-text studies.
Electronic Database	Scopus, IEEE Xplore, Google Scholar	Same studies from different database

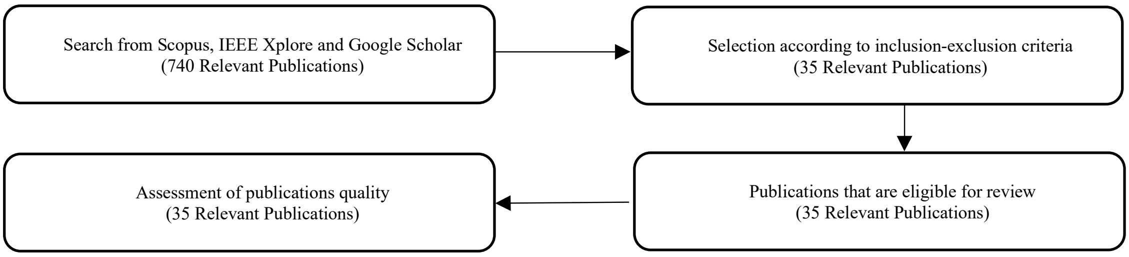

Search from Scopus, IEEE Xplore and Google Scholar
(740 Relevant Publications)
Selection according to inclusion-exclusion criteria
(35 Relevant Publications)
Publications that are eligible for review
(35 Relevant Publications)
Assessment of publications quality
(35 Relevant Publications)

Fig. 1 Illustration Search for relevant publications

The search query returned a total of 740 studies that corresponded to the specified search phrases. The sources of studies included book reports, scholarly journals, academic conferences, theses, or dissertations. Following the application of predetermined inclusion and exclusion criteria, 705 studies did not meet the requirements for inclusion in the analysis. This led to the selection of 35 studies that satisfied the predetermined criteria, comprising 7 conferences, 27 journals, and 1 thesis. Fig. 1 shows an exemplification of the process included in conducting this study. Details of the selected studies for further review are presented in Table 4 as types of publications and datasets used.

Before the review process, a comprehensive evaluation was conducted to assess the quality of all studies. The enhanced analysis method exclusively incorporated studies of suitable quality. In this study, the quality analysis method used was derived from the investigation conducted by [15][16]. Table 5 shows the five categories used for quality assessment, consisting of three distinct value ranges of response, where -1 corresponds to the response "no," 0 indicates "sufficient/partially," and 1 signifies "yes".

TABLE 4
TYPES OF PUBLICATIONS AND DATASETS USED

Author	Language	Dataset		Types of Publications
		Basic Word	Affixed Test Word	
[19]	Ngoko Javanese	Not Stated	Not Stated	Journal
[20]	Javanese	Not Stated	366	Journal
[21]	Javanese	300	300	Journal
[22]	Javanese	Not Stated	Not Stated	Journal
[23]	Javanese	Not Stated	449	Journal
[24]	Javanese	1511	430	Journal Scopus
[25]	Krama Alus Javanese	Not Stated	41	Conference
[26]	Javanese	Not Stated	13011	Conference
[27]	Javanese	3180	1294	Conference
[28]	Balinese	376	376	Journal
[29]	Balinese	Not Stated	357	Journal
[30]	Balinese	10000	2250	Journal
[31]	Balinese	Not Stated	Not Stated	Journal
[32]	Balinese	1000	10 queries	Journal
[33]	Balinese	376	376	Journal
[34]	Sundanese	Not Stated	Not Stated	Conference
[35]	Sundanese	3463	494	Journal
[36]	Sundanese	Not Stated	4453	Journal Scopus
[37]	Sundanese	7871	2945	Journal
[38]	Sundanese	11885	546	Conference
[39]	Madurese	Not Stated	400	Journal
[40]	Madurese	17055	50 Sentences	Conference
[41]	Madurese	2259	1000	Journal
[42]	Madurese	2295	1000	Journal
[43]	Rejang	6983	9000	Journal
[44]	Rejang	Not Stated	Not Stated	Journal Scopus
[45]	Angkola Batak	Not Stated	782	Journal Scopus
[46]	Angkola Batak	3608	450	Journal
[47]	Minangkabau	600	600	Conference
[48]	Minangkabau	600	600	Journal Scopus
[49]	Riau Malay	3587	1000	Journal
[50]	Kaili	4127	1000	Thesis
[51]	Lampung Api	2000	Not Stated	Journal
[52]	Tetun	176	211	Journal
[53]	Sasak	Not Stated	200	Journal

Based on the assessment results presented in Fig. 2, the evaluation achieved a perfect score of 100% for the first criterion. The assessment showed that 91.43% of the studies comprehensively explained the proposed solution for the second criterion. The third criterion assessment had a significant value of 94.29% in evaluating and validating the proposed method. The fourth criterion scores 100%, showing that all publications are derived from studies rather than the author's viewpoint. Consequently, this study lacks subjective explanations but provides objective explanations supported by experimental data. The results from the evaluation of the fifth criterion show that subsequent studies cited 60% under consideration. This can be attributed to the occurrence of multiple new publications in 2023. Quality analysis shows that all pertinent studies are satisfactory and merit inclusion in subsequent stages of examination.

TABLE 5
QUALITY ASSESSMENT FORM

Item	Assessment Criteria	Score	Description
QA1	Was the objective of the study well articulated?	-1	No
		0	Partially
		1	Yes
QA2	Does the study provide a comprehensive description of the recommended approach?	-1	No
		0	Partially
		1	Yes
QA3	Has the suggested method been empirically validated?	-1	No
		0	Partially
		1	Yes
QA4	Does the study present a specific position or opinion?	-1	No
		0	Partially
		1	Yes
QA5	Do other academics cite or reference the study?	-1	No
		0	Partially
		1	Yes

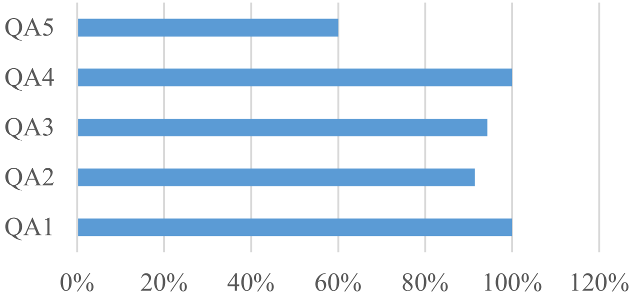


Fig. 2 Quality assessment result from 35 studies

B. Reporting the Results of the Literature Review

The results are presented through a comprehensive analysis of the study summary and by addressing each question individually. The characterization of each question is derived from the results obtained during data extraction. Although previous studies [6] [7] have investigated SLR for stemming, there is a need for specific practices of stemming and lemmatization applied to regional languages in Indonesia.

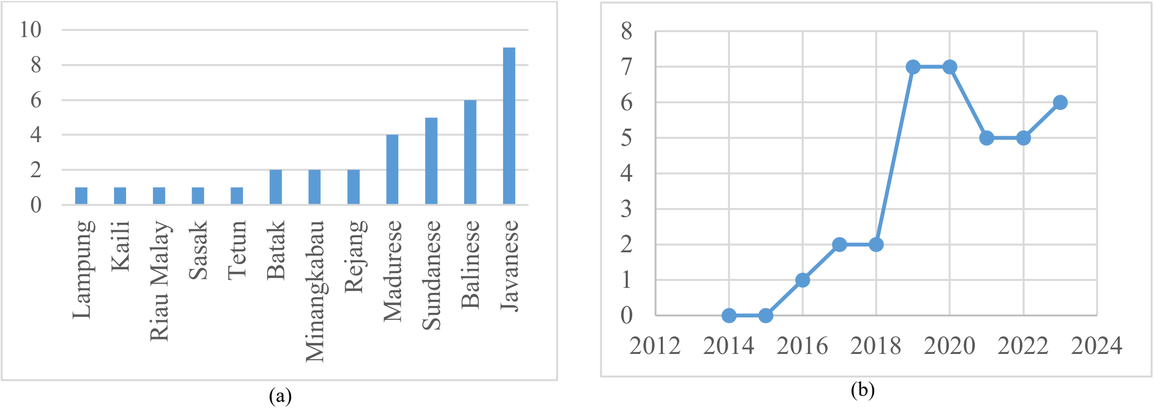
IV. Results

A. Summary of Studies Relevant to the Research Question

The results of the literature review process show that the number of relevant publications for further study according to the inclusion-exclusion criteria is 35. Fig. 3 shows the number of studies by regional languages and year of publication, while Table 6 presents the distribution of the 35 relevant publications, consisting of 27 academic journals, 7 national or international conferences, and 1 thesis.

TABLE 6
Distribution of 35 Relevant Publications

Type of Publication	Number of Publications
Journal	27
Proceeding	7
Thesis	1



(a)

(b)

Fig. 3 Visualization of 35 relevant publications based on (a) regional languages; (b) number of studies each year

TABLE 7
LITERATURE REVIEW OF STEMMING METHODS AND SOURCE REFERENCES

Author	Language	Method	Source
[19]	Ngoko Javanese	Modification of Enhanced Confix Stripping Stemmer	Google Scholar
[20]	Javanese	Modification of Nazief and Adriani	Google Scholar
[21]	Javanese	Damerau Levenshtein Distance	Google Scholar
[22]	Javanese	Rule-based using Morphology	Google Scholar
[23]	Javanese	Enhanced Confix Stripping Stemmer for Modification of Nazief and Adriani	Google Scholar
[24]	Javanese	Rule-based using Morphology and String Matching	Scopus
[25]	Krama Alus Javanese	Modification of Nazief and Adriani	IEEE Xplore
[26]	Javanese	Modification of Nazief and Adriani	IEEE Xplore
[27]	Javanese	Transformer	IEEE Xplore
[28]	Balinese	Rule-based using Morphology	Google Scholar
[29]	Balinese	Modification of the Porter Stemmer Algorithm	Google Scholar
[30]	Balinese	Levenshtein Distance	Google Scholar
[31]	Balinese	Modification of Nazief and Adriani	Google Scholar
[32]	Balinese	Rule-based using Morphology and N-Gram	Google Scholar
[33]	Balinese	Enhanced Confix Stripping Stemmer	Google Scholar
[34]	Sundanese	Not stated	IEEE Xplore
[35]	Sundanese	Syllable Pattern/Canonical-based	Google Scholar
[36]	Sundanese	Rule-based using Morphology	Scopus
[37]	Sundanese	Multi Rule-based and Corpus-based	Google Scholar
[38]	Sundanese	Sundanese Stemmer Based on Morphophonemics	IEEE Xplore
[39]	Madurese	Modification of Enhanced Confix Stripping Stemmer	Google Scholar
[40]	Madurese	Modification of Enhanced Confix Stripping Stemmer combined with the concept of Rule-Based word morphology	IEEE Xplore
[41]	Madurese	Modification of Nazief and Adriani & Modification of Enhanced Confix Stripping Stemmer	Google Scholar
[42]	Madurese	Modification of Nazief and Adriani	Google Scholar
[43]	Rejang	Modification of Enhanced Confix Stripping Stemmer, New Enhanced Confix Stripping Stemmer	Google Scholar
[44]	Rejang	Enhanced Confix Stripping Stemmer	Scopus
[45]	Angkola Batak	Rule-based using Morphology	Scopus
[46]	Angkola Batak	Rule-based using Morphology	Google Scholar
[47]	Minangkabau	Rule-based using Morphology	IEEE Xplore
[48]	Minangkabau	Rule-based using Morphology	Scopus
[49]	Riau Malay	Rule-based using Morphology	Google Scholar
[50]	Kaili	Modification of Nazief and Adriani	Google Scholar
[51]	Lampung Api	Brute Force/Table Look Up	Google Scholar
[52]	Tetun	Rule-based using Morphology	Google Scholar
[53]	Sasak	Modification of the Porter Stemmer Algorithm	Google Scholar

This study successfully obtained factual information to address Research Questions 1, 2, and 3, based on insights from a comprehensive analysis of 35 studies. Fig. 3 and Table 6 summarize the progression of studies on stemming and lemmatization in regional languages from 2014 to 2023. These studies predominantly used the affix removal method and digital dictionaries. The primary objective was to investigate the morphological aspects of a specific regional language and construct a digital lexicon. After reading 35 studies, it was discovered that only 34 focused on stemming and 1 emphasized lemmatization.

Table 7 presents data on publications regarding regional languages and reference source details. Meanwhile, Table 8 provides information on publications in local languages, including the year and objective of the study.

B. What are the advantages and disadvantages of published studies on stemming and lemmatization of Indonesian regional languages?

This section presents and discusses detailed results regarding the advantages and disadvantages of 35 relevant publications. The results include stemming and lemmatization of Javanese, Balinese, Sundanese, Madurese, Rejang, Batak Angkola, Minangkabau, Riau Malay, Kaili, Lampung, Tetun, and Sasak languages. The study has examined all publications on stemming or lemmatization in regional languages in Indonesia and evaluated their strengths and weaknesses. This was carried out by examining the information available in the introduction, methods, results, discussion, and references sections. The introduction section generally explained the background of the problem (X1), the motivation for the study (X2), objectives (X3), and contributions (X4). The literature review section identifies relevant previous studies (X5) and juxtaposes with the proposed method (X6). The methodology section describes in detail the methods (X7), models or algorithms used in the study (X8), and data availability (X9). The results section describes the outcomes, advantages, and disadvantages of the proposed method (X10) and compares with other methods (X11). The discussion section addresses matters other than the results (X12), and the conclusion shows the achievement of the study objectives and other results (X13), including opportunities for future work (X14). The reference section reviews whether the sources used are valid or not (X15).

TABLE 8
LITERATURE REVIEW OF STEMMING METHODS AND YEAR, OBJECTIVE

Author	Language	Year	Objective
[19]	Ngoko Javanese	2023	This study investigates the use of Enhanced Confix Stripping (ECS) in Ngoko Javanese stemmer.
[20]	Javanese	2020	Nazief and Adriani algorithm stems from Javanese-influenced terms in this investigation.
[21]	Javanese	2021	This work uses the Damerau Levenshtein Distance algorithm, a string-matching method, to identify the fundamental word structures of Javanese.
[22]	Javanese	2018	This study aims to implement a Javanese language stemmer with the rule-based method.
[23]	Javanese	2019	The objective of this study is to enhance the efficiency of Nazief and Adriani alterations by implementing the Enhanced Confix Stripping (ECS) modification method.
[24]	Javanese	2017	This study presents a method for deriving the root form of the Ngoko Javanese language.
[25]	Krama Alus Javanese	2019	Krama Alus Javanese news study group offers viewers the ability to search for news based on categories using the Hierarchical and K-Means Algorithms.
[26]	Javanese	2020	Stemming algorithm in this study is derived by modifying Nazief and Adriani algorithms.
[27]	Javanese	2021	For Transformer encoder-decoder affixes, affix characters as a unit was proposed. This corpus includes affixed, canonical affixed, and non-affixed Javanese words.
[28]	Balinese	2020	In this study, Balinese stemming process uses the Rule Base method.
[29]	Balinese	2017	This study aims to modify the Porter Stemmer algorithm on Balinese affixes.
[30]	Balinese	2020	This study aims to overcome the shortcomings of the ruled-based method with the Levenshtein distance method.
[31]	Balinese	2019	In this study, the Bastal method is used to stem Balinese text to find the basic word. The Bastal algorithm adapts Nazief & Adriani.
[32]	Balinese	2019	This study aims to combine rule-based and n-gram methods for stemming Balinese words.
[33]	Balinese	2020	This study aims to assess the effectiveness of the Enhanced Confix Stripping Stemmer (ECS) algorithm on Balinese stemming
[34]	Sundanese	2020	This study examines the emotion classification of Sundanese text. The preprocessing stage includes several steps, namely case folding, stopword removal, stemming, tokenization, and text representation.
[35]	Sundanese	2021	This project will improve stemming with a canonical syllable pattern based on Sundanese phonology.
[36]	Sundanese	2018	This study aims to build a rule-based stemmer of Sundanese language
[37]	Sundanese	2022	This work aims to build stemming method by comparing the outcomes of morphological (affix and pro-lexeme removal), syllable (canonical), and corpus data.
[38]	Sundanese	2023	This work aims to enhance earlier studies by combining Stemming Process into a rule-based language processing algorithm, called AM Sundanese Stemmer (AMS).
[39]	Madurese	2016	This study modified the Enhanced Confix Stripping Stemmer (ECS) algorithm on Madurese language.
[40]	Madurese	2022	This study modified the Enhanced Confix Stripping Stemmer (ECS) algorithm on Madurese language combined with the concept of Rule-Based word morphology.
[41]	Madurese	2023	This study aims to compare the modified ECS and Nazief and Adriani algorithms on Madurese language.
[42]	Madurese	2023	This study aims to compare the modified Nazief and Adriani algorithms on Madurese language.
[43]	Rejang	2022	The utilized algorithms are Enhanced Confix Striping (ECS), New Enhanced Confix Striping (NECS), and Rejang Algorithm.
[44]	Rejang	2019	To construct the Rejang language morphology algorithm, the Indonesian UG18 algorithm was studied and analyzed for its strengths and faults.
[45]	Angkola Batak	2019	To separate keywords in the Angola-Mandailing Batak language document, which has different phonologies and morphologies from Indonesian, the modified Confix stripping stemmer method is used.
[46]	Angkola Batak	2023	This study examines and formulates stemming method for Batak Angkola language using grammatical principles.
[47]	Minangkabau	2022	This study aims to create stemming-based Minangkabau language translation machine.
[48]	Minangkabau	2023	The objective of this work is to construct a translation system for ML into Indonesian through the advancement of the NLP idea.
[49]	Riau Malay	2021	In this study, analyze and design Riau Malay stemming algorithm based on morphological rules.
[50]	Kaili	2023	Stemming algorithm in this study is derived by modifying Nazief and Adriani algorithm.
[51]	Lampung Api	2021	This study aims to use the Brute Force method for stemming Lampung Api language.
[52]	Tetun	2019	This study aims to create a ruled-based algorithm for Tetun language.
[53]	Sasak	2019	This study aims to modify the porter stemmer algorithm for Sasak language.

The advantage of the Shevia et al. [19] is that using ECS can increase stemming accuracy from 62% to 97% compared to ruled-based. However, the limitations are based on the accuracy of the results obtained, which are still below 100%, showing the need for a deeper study of ECS rule method. The advantage of Aji et al. [20] is that the use of Nazief and Adriani modification obtains stemming accuracy of 95.9%. Meanwhile, the disadvantage is that the accuracy obtained is still below 100%, showing the need for further investigation on Nazief and Adriani rule, specifically for the suffix and infix. The study by Aji et al. [21] uses a string matching algorithm, namely Damerau Levenshtein Distance. Meanwhile, the disadvantage is the measurement of only the distance between affixed words and the base word, without considering ignores the morphological aspects. Fatkhul and Jefri [22] offered the advantage of using rule-based stemming, adapted from the Porter Stemmer algorithm in Javanese, which is applied to a Javanese language text document information retrieval system. The 77% accuracy value lacks an explanation, despite the use of primary word data, test words, and digital dictionaries. Another study by Nur et al. [23] successfully modified ECS in Javanese to overcome the shortcomings of Nazief and Adriani method in increasing the accuracy value of stemming. The weakness is that ECS has not been able to stem the words *ngetan*, *kumanggah*, *kumarut*, *kumasis*, *kumareg*, *kumadul*, *kumaras*, *katawakake*, and *pangenan*. Fatkhul et al. [24] used ruled-based and string matching for Ngoko Javanese stemming, but was limited to prefixes and suffixes. This shows the inability of the method to overcome infixes, confixes, and repeated words. The strength of the study by Denis et al. [25] lies in its thorough implementation of Javanese stemming using Nazief and Adriani method for Krama Alus Javanese news clustering task. However, a drawback is that the accuracy of stemming is only 75.6%, showing the need for improvement. The application of Javanese stemming from Nazief and Adriani method to prefix, suffix, infix, confix, and repeated words is the advantage of Mohammad et al. [26]. The limitation of the results is the imperfection rules and incomplete dictionaries, negatively impacting the accuracy of stemming results. Sri et al. [27] successfully use affix characters as units in the transformer architecture for morphological segmentation of Javanese words. The results are limited based on accuracy level below 81%, showing the importance of further investigation.

Putu et al. [28] applied rule-based Balinese language stemming from Balinese text documents. However, the study excludes confixes and bases stemming from Balinese words, which had a high affix complexity. The advantage of Gusti et al.'s publication [29] is the ability to stem Balinese language using a modified Porter stemmer, which uses Balinese morphology. This stemmer is applied to translate Balinese documents into Indonesian. Despite the advantage, the method only performs stemming on ECS prefixes and suffixes, which can interfere with the translation of Balinese language documents. Gede et al. [30] successfully applied lemmatization of Balinese language, using a rule and lexicon-based method, conducting experiments with and without Levenshtein distance. The weakness of the results is the incorrect use of the term "lemmatization," as Balinese language lacks distinctions between first, second, and third verbs, compared to English. The publication of Putu et al. [31] also uses the application of Nazief and Adriani stemming modification to Balinese to Indonesian language translation machine. Meanwhile, the limitation is that the accuracy of sentence translation reaches 86.67%, showing the need for improvements in rule. Made and Chatine [32] successfully applied a combination of rule-based and N-Gram stemming for Indonesian Balinese, as well as conducted tests using sentences as queries. The weakness is that a semantic aspect is required in the sentence-shaped test query. Ni et al. [33] showed that Balinese stemming uses ECS to improve previous studies [28], increasing accuracy by 19.12%, although ECS method could not stem particular prefix decays.

The application of Sundanese stemming to the emotion classification task is properly detailed by Oddy et al. [34], but the results lack an explanation of stemming method. Ade et al. [35] has the advantage of using a distinct stemming method, namely the Syllable Pattern method, but accuracy only reaches 89%, showing the need for improvement. Arie et al. [36] shows that Sundanese stemming uses a rule-based method, with the baseline being the confix stripping method. However, the weakness of this rule-based method is the inability to stem words that are contextually ambiguous and reduplication. Ade et al. [37] showed that Sundanese language stemming uses two different methods, namely the ruled-based and the corpus-based. The weakness of these methods is the need to stem compound and abbreviation words. Aries et al. [38] successfully conducted Sundanese stemming experiments with three methods, namely Nazief and Adriani modification, UG 18 modification, and AM Sundanese stemmer. The best results were obtained by AM Sundanese stemmer regarding accuracy, recall, precision, and F1-score. The weakness is that modifying the AM Sundanese Stemmer algorithm is essential, as the accuracy is still below 100%.

Rakhmad [39] shows the availability of Madurese morphology, although there is no explanation of ECS algorithm used or information on the accuracy value obtained. The application of ECS stemming from the Madurese language translation machine is an advantage of Fika et al. [40]. However, the results showed accuracy value below 90%, indicating the need for improvement in the modification of ECS stemming in Madurese. Enni et al.'s publications [41] [42] successfully used Nazief and Adriani and ECS algorithms to stem the Madurese language. However, the first publication fails to stem infix and reduplication [41], including the second report [42].

Sastya et al. [43] is the only stemming study that discusses the complexity of algorithm, using three methods, namely ECS, New ECS, and Rejang Algorithm. However, there is a need to modify the algorithm due to the inability to recognize 27 among the 9000 words provided. Another study conducted by Sastya et al. [44] showed details of the morphology of the Rejang language along with each flowchart. The result did not consider the test and base word dictionary used or the accuracy achieved. M A et al. [45] successfully developed ECS-based stemming algorithm for the Angkola Batak language comprehensively by testing using documents. The weakness is that some prefix rules fail to be recognized, showing the need for detailed morphological studies. The strength of publication by Nur et al. [46] is the development of a rule-based stemming algorithm based on the morphology of the Angkola Batak language comprehensively. The weakness is the use of a less competitive digital dictionary, with an accuracy of 99.56%. Rini [47] and [48] discussed the morphology of Minangkabau language and their application to the translation of Minangkabau documents. However, the results obtained accuracy below 100%, showing the need for method improvement.

Yusra et al. [49] successfully developed a rule-based stemming algorithm based on the morphology of Riau Malay language comprehensively, obtaining an accuracy of 100%. Tamrizal [50] also developed 24 modified Nazief and Adriani algorithms for Kaili language in Central Sulawesi, but the accuracy was 93–96%. The study by Zaenal et al. [51] showed the use of the Brute-Force method, providing a link to the list of test words. However, the weakness is the inability to dissect affixed words in the Lampung language. Anita et al. [52] comprehensively explained Tetun language morphology, but the reports did not include the results section. Additionally, there was no placement of Tetun language stemming experiment results in the method/system design section and only three references were provided. Yulita et al. [53] is the only stemming study that uses the Porter Stemmer method for the Sasak language in Lombok. The weakness is the need for more explanation in the Porter-Stemmer method section for the Sasak language.

C. What methods do text stemming or text lemmatization studies apply in Indonesia for texts originating from regional languages?

Previous studies conducted in Indonesia have used various methods to address text stemming and lemmatization issues in regional languages. These include (1) the development of rule-based and statistical hybrid algorithms to reduce several limitations of each method. (2) Other academics use machine learning to construct regional language stemming and lemmatization methods. This strategy could be effective in languages with complex morphology and several meaning interpretations. (3) Recent studies are also building corpora and annotated data sets for languages to train and test language-specific stemming and lemmatization algorithms. (4) Resources and information are being shared to accelerate Indonesian regional languages stemming and lemmatization algorithms.

A manually designed linguistic rule-based stemming method strips words of "-s" and reduces to their root. This method works well in English and is easy to understand, but requires time to build rules and may struggle with irregular words or be difficult to use in other languages. Morphology-based stemming uses linguistic understanding to identify and remove prefixes, suffixes, and roots. Dictionaries or morphological analysis help grasp word meanings and connections. It also handles foreign languages and complicated patterns better than rule-based systems, which are computationally expensive and language-intensive. Similar to guessing, brute force stemming rejects word components to obtain the root. Table look-up uses a pre-existing reference list to promptly respond to only the specified elements. In comparison, brute force method distorts words, while table look-up misses uncommon terms.

TABLE 9
METHODS FOR STEMMING AND LEMMATIZATION IN INDONESIAN REGIONAL LANGUAGES

Author	Language	Method
[20]	Javanese	Modification of Nazief and Adriani
[25]	Krama Alus Javanese	
[26]	Javanese	
[31]	Balinese	
[41]	Madurese	
[42]		
[50]	Kaili	
[19]	Ngoko Javanese	Modification of Enhanced Confix Stripping Stemmer
[23]	Javanese	
[33]	Balinese	
[39]	Madurese	
[40]		
[43]	Rejang	
[44]		
[22]	Javanese	Rule-Based using Morphology
[24]		
[28]	Balinese	
[32]		
[36]	Sundanese	
[37]		
[38]		
[45]	Angkola Batak	
[46]		
[47]	Minangkabau	
[48]		
[49]	Riau Malay	
[52]	Tetun	
[29]	Balinese	Modification of the Porter Stemmer Algorithm
[53]	Sasak	
[21]	Javanese	Damerau Levenshtein Distance
[27]		Transformer
[30]	Balinese	Levenshtein Distance
[34]	Sundanese	Not stated
[35]		Syllable Pattern/Canonical-based
[51]	Lampung Api	Brute Force/Table Look Up

Table 9 provides a summary of the results from the evaluation of the 35 publications selected for SLR. Specifically, it focuses on the various methods used by studies conducted in Indonesia to create algorithms for stemming, stemmer, lemmatization, and lemmatizer for regional languages.

A Bahasa Indonesia specialist who removes common affixes and examines word structure, particularly in difficult cases, uses a method similar to the modified Nazief and Adriani. This method handles complex morphology and overcomes stemming constraints, showing suitability for regional languages. In natural language processing, stemming includes a modified ECS method to simplify words to their roots. Compared to other stemming algorithms, modified Nazief and Adriani consider prefixes and suffixes. In complex morphology languages, prefixes and suffixes improve stemming accuracy. Word stems are found through systematic deletion, as well as consideration of language-specific rules and exceptions, thereby improving regional language comprehension, text mining, and information retrieval.

D. What are the challenges of text stemming and lemmatization in regional languages in Indonesia?

Text stemming and lemmatization in Indonesian regional languages have several challenges. (1) Indonesia has over 700 regional languages, each with distinct morphology and syntax, thereby limiting the development of a single algorithm to stem and lemmatize words in all these languages. (2) Scarce resources due to the lack of annotated regional languages text corpora containing morphological and syntactic components is a problem. This makes training and evaluating stemming and lemmatization algorithms for different languages difficult. (3) Indonesian regional languages have complicated prefixes, suffixes, and infixes, resulting in difficult determination of word-based form. (4) Some Indonesian regional languages have polysemous vocabulary words, with meanings altered by context, leading to the inability to determine the right lemma. (5) The lack of standardized spelling contributes to the challenge of searching for the right word form for stemming and lemmatization.

Despite the challenges, several significant advancements have been made in stemming and lemmatization algorithms for regional languages in Indonesia. These include the development of tools and libraries capable of stemming and lemmatizing text. The tools have been used in several practical contexts, including machine translation, text classification, and information retrieval. For future studies, significant progress will be made in enhancing text stemming and lemmatization algorithms for regional languages in Indonesia. This is due to the increasing availability of resources and the growing number of studies dedicated to addressing the challenges.

V. Discussion

The development of text stemming and lemmatization studies in regional languages in Indonesia experienced several challenges between 2014 - 2023. During the 2014–2020 period presented in Fig. 2, the trend of text stemming and lemmatization studies increased. This is followed by a significant decrease in 2021–2023, but there is potential for a rise. Specifically, the number of stemming and lemmatization studies is still dominated by Javanese with 9 publications [19][20][21][22][23][24][25][26][27], 6 Balinese [28][29][30][31][32][33], 5 Sundanese [34][35][36][37][38], and 4 Madurese [39][40][41] [42], while the regional languages only have 2 or 1 publication.

Several factors play a significant role in shaping the progress on stemming or lemmatization. These include (1) resource availability, as the successful development of stemming and lemmatization algorithms requires access to various resources, including text corpora, annotated data sets, and computational resources. The accessibility of these resources can substantially influence the progress rate in this field of study. The availability of digital dictionaries, structured explanations through word morphology books in specific regional languages, and regional language experts are also essential. (2) The community's interest also influences the development of stemming and lemmatization studies. When a significant degree of interest is present, more studies will be carried out on this issue, potentially accelerating the pace of advancement. Technological improvements can have a favorable influence in the fields of stemming and lemmatization. (3) The practical use of stemming and lemmatization algorithms includes diverse applications to machine translation, text classification, and information retrieval. The occurrence of novel applications for these algorithms has the potential to generate significant attention and enthusiasm related to stemming and lemmatization.

The advancement of study on stemming and lemmatization might be influenced by particular requirements. For example, the development of a novel language requires the creation of specific algorithms for stemming and lemmatization. Additionally, the development of new applications requires the creation of unique algorithms focusing on meeting specified needs. Through a thorough examination of these factors influencing the development of algorithms, a more complete understanding can be obtained to facilitate the discovery of possible study directions.

The limitation of this SLR study is that the analysis focuses on studies related to stemming and lemmatization in regional languages in Indonesia. Furthermore, only studies indexed by Scopus, IEEE Xplore, and Google Scholar from 2014 - 2023 are included, without considering other local language stemming studies in the form of undergraduate theses.

VI. Conclusion

In conclusion, this SLR stemming study successfully provided information on the development of stemming and lemmatization studies in 2014-2023. The results showed the strategies used and potential challenges encountered in other regional languages in Indonesia. Among 35 studies selected, Javanese, Balinese, Madurese, and Sundanese were dominant topics for stemming, which reported quantitative results. For regional languages that would undertake stemming or lemmatization studies, the presence of a digital dictionary was the main supporting factor, particularly when using various rule-based methods, including detailed information related to morphological studies of the regional languages. Furthermore, the greater number of essential words in the digital dictionary correlated with more positive contributions to stemming and lemmatization accuracy. A corpus-based stemming method was only possible when there were several corpora available to represent the variety of affix words in languages. This SLR provided an overview of the sequence of methods used, starting from rule-based, modification of Nazief & Adriani algorithm, and modification of ECS Stemmer.

Author Contributions: *Zaenal Abidin*: Conceptualization, Methodology, Writing - Original Draft, Writing. *Akmal Junaidi* : Methodology, Writing – Review & Editing, Supervision. *Wamiliana*: Methodology, Writing – Review & Editing, Supervision.

All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by Direktorat Jenderal Riset, Teknologi dan Pengabdian Kepada Masyarakat (DRTPM), Ministry of Education, Culture, Research, and Technology, Republic of Indonesia, under the scheme of Doctoral Research (PDD) with contract number 131/E5/PG.02.00.PL/2023.

Acknowledgments: We thank Ministry of Education, Culture, Research, and Technology, Republic of Indonesia, for supporting this research.

Conflicts of Interest: The authors declare no conflict of interest.

Data Availability: Data sharing is not applicable to this article as no new data were created or analyzed in this study.

Institutional Review Board Statement: Not applicable.

Informed Consent: There were no human subjects.

Animal Subjects: There were no animal subjects.

ORCID:
Zaenal Abidin: https://orcid.org/0000-0003-4237-7167
Akmal Junaidi: https://orcid.org/0000-0003-1030-6954
Wamiliana: https://orcid.org/0000-0002-3740-7950

References

[1] J. Singh and V. Gupta, “A systematic review of text stemming techniques,” *Artif Intell Rev*, vol. 48, no. 2, pp. 157–217, Aug. 2017, doi: 10.1007/s10462-016-9498-2.

[2] J. Singh and V. Gupta, “Text stemming: Approaches, applications, and challenges,” *ACM Comput Surv*, vol. 49, no. 3, Sep. 2016, doi: 10.1145/2975608.

[3] A. Jabbar, S. Iqbal, M. I. Tamimy, S. Hussain, and A. Akhunzada, “Empirical evaluation and study of text stemming algorithms,” *Artif Intell Rev*, vol. 53, no. 8, pp. 5559–5588, Dec. 2020, doi: 10.1007/s10462-020-09828-3.

[4] A. S. Rizki, A. Tjahyanto, and R. Trialih, “Comparison of stemming algorithms on Indonesian text processing,” *Telkomnika (Telecommunication Computing Electronics and Control)*, vol. 17, no. 1, pp. 95–102, Feb. 2019, doi: 10.12928/TELKOMNIKA.v17i1.10183.

[5] A. Jabbar, S. Iqbal, M. I. Tamimy, A. Rehman, S. A. Bahaj, and T. Saba, "An Analytical Analysis of Text Stemming Methodologies in Information Retrieval and Natural Language Processing Systems," *IEEE Access*, vol. 11, pp. 133681–133702, 2023, doi: 10.1109/ACCESS.2023.3332710.

[6] R. Pramana, Debora, J. J. Subroto, A. A. S. Gunawan, and Anderies, "Systematic Literature Review of Stemming and Lemmatization Performance for Sentence Similarity," in *Proceedings of the 2022 IEEE 7th International Conference on Information Technology and Digital Applications, ICITDA 2022*, Institute of Electrical and Electronics Engineers Inc., 2022. doi: 10.1109/ICITDA55840.2022.9971451.

[7] A. Maesya, A. Ramadhan, E. Abdurachman, A. Trisetyarso, and M. Zarlis, "Stemming Algorithm for the Indonesian Language: A Scientometric View," in *2022 IEEE Creative Communication and Innovative Technology, ICCIT 2022*, Institute of Electrical and Electronics Engineers Inc., 2022. doi: 10.1109/ICCIT55355.2022.10119050.

[8] Y. K. Paskahningrum, E. Utami, and A. Yaqin, "A Systematic Literature Review of Stemming in Non-Formal Indonesian Language," in *International Journal of Innovative Science and Research Technology*, vol. 8, no.1, pp. 62-69, 2023. [Online]. Available: www.ijisrt.com

[9] A. F. Aji *et al.*, "One Country, 700+ Languages: NLP Challenges for Underrepresented Languages and Dialects in Indonesia," pp. 7226–7249, 2022, doi: 10.18653/v1/2022.acl-long.500.

[10] M. D. Sanjaya, B. Indonesia, and D. Daerah, "Bahasa Indonesia dan Daerah Sebagai Perekat Jati Diri dan Martabat Bangsa Di Era Globalisasi" *Jurnal Bindo Sastra*, vol. 1, no. 1, pp. 10-14, 2017.

[11] M. Abstrak, "Eksistensi Bahasa Indonesia Sebagai Bahasa Persatuan." *Jurnal Sosial Humaniora (JSH)*, vol. 1, no.2, pp. 172-184, 2008.

[12] S. Cahyawijaya *et al.*, "NusaCrowd: Open Source Initiative for Indonesian NLP Resources," 2022, [Online]. Available: http://arxiv.org/abs/2212.09648

[13] G. I. Winata *et al.*, "NusaX: Multilingual Parallel Sentiment Dataset for 10 Indonesian Local Languages," *EACL 2023 - 17th Conf. Eur. Chapter Assoc. Comput. Linguist. Proc. Conf.*, pp. 815–834, 2023.

[14] D. Khurana, A. Koli, K. Khatter, and S. Singh, "Natural language processing: state of the art, current trends and challenges," *Multimed Tools Appl*, vol. 82, no. 3, pp. 3713–3744, Jan. 2023, doi: 10.1007/s11042-022-13428-4.

[15] R. Fauzan, D. Siahaan, M. Solekhah, V. W. Saputra, A. E. Bagaskara, and M. I. Karimi, "A Systematic Literature Review of Student Assessment Framework in Software Engineering Courses," *Journal of Information Systems Engineering and Business Intelligence*, vol. 9, no. 2, pp. 264–275, Oct. 2023, doi: 10.20473/jisebi.9.2.264-275.

[16] M. A. W. Putra Rahmadhan, D. I. Sensuse, R. R. Suryono, and Kautsarina, "Trends and Applications of Gamification in E-Commerce: A Systematic Literature Review," *Journal of Information Systems Engineering and Business Intelligence*, vol. 9, no. 1, pp. 28–37, Apr. 2023, doi: 10.20473/jisebi.9.1.28-37.

[17] B. Kitchenham, O. Pearl Brereton, D. Budgen, M. Turner, J. Bailey, and S. Linkman, "Systematic literature reviews in software engineering - A systematic literature review," *Information and Software Technology*, vol. 51, no. 1. pp. 7–15, Jan. 2009. doi: 10.1016/j.infsof.2008.09.009.

[18] B. Kitchenham *et al.*, "Systematic literature reviews in software engineering-A tertiary study," *Information and Software Technology*, vol. 52, no. 8. Elsevier B.V., pp. 792–805, 2010. doi: 10.1016/j.infsof.2010.03.006.

[19] S. I. Melia, J. Sholihah, D. Nisak, I. S. Juniaristha, and A. T. Ni'mah, "The Ngoko Javanese Stemmer uses the Enhanced Confix Stripping Stemmer Method," *Rekayasa*, vol. 16, no. 1, pp. 107–112, Apr. 2023, doi: 10.21107/rekayasa.v16i1.19308.

[20] A. P. Wibawa, F. A. Dwiyanto, I. A. E. Zaeni, R. K. Nurrohman, and A. Afandi, "Stemming javanese affix words using nazief and adriani modifications," *Jurnal Informatika*, vol. 14, no. 1, p. 36, Jan. 2020, doi: 10.26555/jifo.v14i1.a17106.

[21] A. P. Wibawa and M. N. Hakim, "Stemming Bahasa Jawa Menggunakan Damerau Levenshtein Distance (Dld)," *J. Tek. Inform.*, vol. 14, no. 1, pp. 22–27, 2021, doi: 10.15408/jti.v14i1.15010.

[22] F. Amin and J. Alfa Razaq, "Implementasi Stemmer Bahasa Jawa dengan Metode Rule Base Approach pada Sistem Temu Kembali Informasi Dokumen Teks Berbahasa Jawa," *Pros. SENDI_U*, pp. 199–206, 2018.

[23] N. Hidayatullah, Aji Prasetya Wibawa, and Harits Ar Rosyid, "Penerapan ECS Stemmer untuk Modifikasi Nazief & Adriani Berbahasa Jawa," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 3, no. 3, pp. 343–348, 2019, doi: 10.29207/resti.v3i3.994.

[24] F. Amin, W. Hadikurniawati, S. Wibisono, H. Februariyanti, and J. S. Wibowo, "A hybrid method of rule-based and string matching stemmer for Javanese language," *J. Theor. Appl. Inf. Technol.*, vol. 95, no. 19, pp. 4973–4982, 2017.

[25] D. E. Cahyani, L. M. T. Utami, and H. Setiadi, "Clustering of Javanese news in Krama Alus Level with Javanese stemming," *2019 Int. Conf. Inf. Commun. Technol. ICOIACT 2019*, pp. 462–467, 2019, doi: 10.1109/ICOIACT46704.2019.8938438.

[26] M. A. Nq, L. P. Manik, and D. Widiyatmoko, "Stemming Javanese: Another Adaptation of the Nazief-Adriani Algorithm," in *2020 3rd International Seminar on Research of Information Technology and Intelligent Systems, ISRITI 2020*, Institute of Electrical and Electronics Engineers Inc., Dec. 2020, pp. 627–631. doi: 10.1109/ISRITI51436.2020.9315420.

[27] S. H. Wijono, M. R. Alhamidi, M. H. Hilman, and W. Jatmiko, "Canonical Segmentation Using Affix Characters as a Unit on Transformer for Javanese Language," in *Proceedings - IWBIS 2021: 6th International Workshop on Big Data and Information Security*, Institute of Electrical and Electronics Engineers Inc., 2021, pp. 67–72. doi: 10.1109/IWBIS53353.2021.9631839.

[28] N. W. Wardani and P. G. S. Cipta Nugraha, "Stemming Dokumen Teks Bahasa Bali Dengan Metode Rule Base Approach," *JATISI (Jurnal Tek. Inform. dan Sist. Informasi)*, vol. 7, no. 3, pp. 510–521, 2020, doi: 10.35957/jatisi.v7i3.538.

[29] G. N. M. Nata and P. P. Yudiastra, "Stemming teks sor-singgih Bahasa Bali," *E-Proceedings KNS&I STIKOM Bali*, pp. 608–612, 2017, [Online]. Available: http://knsi.stikom-bali.ac.id/index.php/eproceedings/article/view/111

[30] I. G. A. Purnajiwa Arimbawa and N. A. Sanjaya ER, "Lemmatization in Balinese Language," *JELIKU (Jurnal Elektron. Ilmu Komput. Udayana)*, vol. 8, no. 3, p. 235, 2020, doi: 10.24843/jlk.2020.v08.i03.p04.

[31] I. P. M. Wirayasa, I. M. A. Wirawan, and I. M. A. Pradnyana, "Algoritma Bastal: Adaptasi Algoritma Nazief & Adriani Untuk Stemming Teks Bahasa Bali," *J. Nas. Pendidik. Tek. Inform.*, vol. 8, no. 1, p. 60, 2019, doi: 10.23887/janapati.v8i1.13500.

[32] M. A. P. Subali and C. Fatichah, "Kombinasi Metode Rule-Based dan N-Gram Stemming untuk Mengenali Stemmer Bahasa Bali," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 6, no. 2, p. 219, 2019, doi: 10.25126/jtiik.2019621105.

[33] N. W. Wardani and P. G. S. C. Nugraha, "Stemming Teks Bahasa Bali dengan Algoritma Enhanced Confix Stripping," *International Journal of Natural Science and Engineering*, vol. 4, no. 3, pp. 103–113, Dec. 2020, doi: 10.23887/ijnse.v4i3.30309.

[34] O. V. Putra, F. M. Wasmanson, T. Harmini, and S. N. Utama, "Sundanese Twitter Dataset for Emotion Classification," in *CENIM 2020 - Proceeding: International Conference on Computer Engineering, Network, and Intelligent Multimedia 2020*, Institute of Electrical and Electronics Engineers Inc., Nov. 2020, pp. 391–395. doi: 10.1109/CENIM51130.2020.9297929.

[35] A. Sutedi, R. Elsen, and M. R. Nasrulloh, "Sundanese Stemming using Syllable Pattern," *Jurnal Online Informatika*, vol. 6, no. 2, p. 218, Dec. 2021, doi: 10.15575/join.v6i2.812.

[36] A. Ardiyanti Suryani, D. Hendratmo Widyantoro, A. Purwarianti, and Y. Sudaryat, "The rule-based sundanese stemmer," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 17, no. 4, Jul. 2018, doi: 10.1145/3195634.

[37] A. Sutedi, M. R. Nasrulloh, and R. Elsen, "Multi Rule-based and Corpus-based for Sundanese Stemmer," *Jurnal Online Informatika*, vol. 7, no. 2, pp. 184–192, Dec. 2022, doi: 10.15575/join.v7i2.846.

[38] A. Maesya, Y. Arifin, A. Zahra, and W. Budiharto, "Development of Sundanese Stemmer Based on Morphophonemics," in *10th International Conference on ICT for Smart Society, ICISS 2023 - Proceeding*, Institute of Electrical and Electronics Engineers Inc., 2023. doi: 10.1109/ICISS59129.2023.10291840.

[39] R. Maulidi, "Modifikasi Metode Enhanced Confix Stripping," *Pros. Semin. Nas. FDI*, no. December, pp. 12–15, 2016. [Online]. Available: https://www.researchgate.net/publication/311532738

[40] F. H. Rachman, N. Ifada, S. Wahyuni, G. D. Ramadani, and A. Pawitra, "ModifiedECS (mECS) Algorithm for Madurese-Indonesian Rule-Based Machine Translation," in *2022 International Conference of Science and Information Technology in Smart Administration, ICSINTESA 2022*, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 51–56. doi: 10.1109/ICSINTESA56431.2022.10041470.

[41] E. Lindrawati, E. Utami, and A. Yaqin, "Comparison of Modified Nazief&Adriani and Modified Enhanced Confix Stripping algorithms for Madurese Language Stemming," *INTENSIF: Jurnal Ilmiah Penelitian dan Penerapan Teknologi Sistem Informasi*, vol. 7, no. 2, pp. 276–289, Aug. 2023, doi: 10.29407/intensif.v7i2.20103.

[42] Enni Lindrawati, Ema Utami, and A. Yaqin, "ANoM STEMMER: Nazief &amp; Andriani Modification for Madurese Stemming," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 7, no. 6, pp. 1341–1347, Dec. 2023, doi: 10.29207/resti.v7i6.5086.

[43] S. H. Wibowo, R. Toyib, M. Muntahanah, and Y. Darnita, "Time complexity in rejang language stemming," *JURNAL INFOTEL*, vol. 14, no. 3, pp. 174–179, Aug. 2022, doi: 10.20895/infotel.v14i3.764.

[44] S. H. Wibowo and S. Wibowo, "Development of Stemming Algorithm for Rejang Language Stemmer Based on Rejang Language Morphology View project Development of Stemming Algorithm for Rejang Language Stemmer Based on Rejang Language Morphology," *Article in Journal of Advanced Research in Dynamical and Control Systems*, vol. 11, 2019, [Online]. Available: https://www.researchgate.net/publication/341307354

[45] M. A. Muchtar *et al.*, "Separation of Basic Words in Angkola Batak Text Documents using Enhanced Confix Stripping Stemmer Case: Mandailing Ethnic," in *IOP Conference Series: Materials Science and Engineering*, Institute of Physics Publishing, Oct. 2019. doi: 10.1088/1757-899X/648/1/012024.

[46] N. H. Hrp, M. Fikry, and Y. Yusra, "Algoritma Stemming Teks Bahasa Batak Angkola Berbasis Aturan Tata Bahasa," *Journal of Computer System and Informatics (JoSYC)*, vol. 4, no. 3, pp. 642–648, May 2023, doi: 10.47065/josyc.v4i3.3458.

[47] R. Sovia, S. Defit, and Yuhandri, "Development of the Minangkabau Local Language Translation Machine Based on Stemming," in *Proceeding - 2022 International Symposium on Information Technology and Digital Innovation: Technology Innovation During Pandemic, ISITDI 2022*, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 195–198. doi: 10.1109/ISITDI55734.2022.9944457.

[48] R. Sovia, S. Defit, Yuhandri, and Sulastri, "Development of natural language processing on morphology-based Minangkabau language stemming algorithm," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 31, no. 1, pp. 542–552, Jul. 2023, doi: 10.11591/ijeecs.v31.i1.pp542-552.

[49] Muhammad Fikry and Yusra Yusra, "Stemmer Bahasa Melayu Riau Berdasarkan Aturan Morfologi," *Semin. Nas. Teknol. Inf. Komun. dan Ind.*, no. November, pp. 118–124, 2021, [Online]. Available: https://ejournal.uin-suska.ac.id/index.php/SNTIKI/article/view/14405

[50] A.M, Tamrizal, "Algoritma Stemming untuk Teks Bahasa Kaili", Magister thesis, Pascasarjana Magister Teknik Informatika, Universitas AMIKOM Yogyakarta, Yogyakarta, 2023.

[51] Z. Abidin, A. Wijaya, and D. Pasha, "Aplikasi Stemming Kata Bahasa Lampung Dialek Api Menggunakan Pendekatan Brute-Force dan Pemograman C#," *Jurnal Media Informatika Budidarma*, vol. 5, no. 1, p. 1, Jan. 2021, doi: 10.30865/mib.v5i1.2483.

[52] A. Guterres, Gunawan, and J. Santoso, "Stemming Bahasa Tetun Menggunakan Pendekatan Rule Based," *Teknika*, vol. 8, no. 2, pp. 142–147, Oct. 2019, doi: 10.34148/teknika.v8i2.224.

[53] Y. F. Andriani, E. Utami, and S. Suwanto, "Modifikasi Algoritma Porter Stemmer Untuk Stemming Bahasa Sasak," *J. Inf. J. Penelit. dan Pengabdi. Masy.*, vol. 5, no. 3, pp. 61–64, 2020, doi: 10.46808/informa.v5i3.147.