JAMBURA JOURNAL OF INFORMATICS
Volume 6, No. 1 | April 2024: 1-13
ISSN 2685-4244, doi: 10.37905/jji.v6i1.24954

Exploring stemming techniques in Ambon Malay languages: A systematic literature review

Vinnesa Patricia Carolina[1], Ema Utami[2], Ainul Yaqin[3]

[1,2]Magister Informatika, Universitas AMIKOM Yogyakarta, Sleman, Indonesia
[3]Fakultas Ilmu Komputer, Universitas AMIKOM Yogyakarta, Sleman, Indonesia

Article history:
Received April 20, 2024
Revised April 26, 2024
Accepted May 1, 2024
Published May 4, 2024

Keywords:
Ambon Malay language
Nazief & Adriani
Stemming algorithm
Systematic literature review
Text processing

ABSTRACT. Stemming in Ambonese poses a significant challenge due to its extensive lexicon, encompassing approximately 127,000 base words as recorded in the Kamus Besar Bahasa Indonesia (Indonesian Dictionary). This complexity arises from the task of extracting base words from those with affixes, necessitating the removal of various affixes such as prefixes, infixes, suffixes, and their combinations. This process greatly influences analytical outcomes. To address this linguistic complexity, several stemming algorithms were developed. These include Nazief & Adriani, Enhanced Confix Stripping, Sastrawi, and Tala, each offering unique techniques to handle stemming complexities in Indonesian. The selection of the appropriate algorithm is crucial for ensuring the accuracy and reliability of the stemming process within the analytical framework. In conducted stemming research, there were variations in methods used. The most frequently used algorithm was Nazief & Adriani, with 17 recorded cases, followed by Enhanced Confix Stripping, with 12 cases. Sastrawi, although less frequent, was used in 4 cases, while Tala appeared in 1 case. This diversity reflects the available choices when selecting a fitting stemming method. However, this may relate to ongoing research projects, funding availability, or other external conditions affecting research production during that period. Consequently, stemming research remains an exciting and relevant topic, with the potential for continued growth and significant contributions to future text processing and linguistic research.

Corresponding Author:
Vinnesa Patricia Carolina,
Universitas AMIKOM Yogyakarta,
Jl. Ring Road Utara, Sleman, 051024, Indonesia.
Email: vinnesa.p.c@students.amikom.ac.id

INTRODUCTION

Ambon is located in eastern Indonesia and boasts numerous cultural assets, including its traditional fabrics, ethnic diversity, and the local language. Language serves as the foundation of culture, with various ethnic groups in Ambon using their respective languages. It is essential to understand that the Ambonese community primarily uses Ambonese. However, in our increasingly interconnected world, communication abilities in minority languages such as Ambonese are often limited.

Ambonese, also known as Ambonese Malay, is one of the dialects spoken by the inhabitants of the Maluku region (Meturan et al., 2023). In practice, this language has absorbed many influences from various languages, such as Makassar's Malay and Portuguese, due to colonization. Additionally, as Indonesian became the standardized language and widely used by the population, Ambonese also absorbed vocabulary from Indonesian, with the lexicon reflecting the local environment or the habits and lifestyles of the Maluku people (Pesiwarissa, 2023).

Stemming separates base words from affixed words in a sentence by isolating base words and affixes, which may consist of prefixes, infixes, and suffixes (Wibowo et al., 2022). Stemming algorithms vary for each language (Tuhpatussania et al., 2022). In Indonesian stemming, several commonly used stemming methods include Nazief-Adriani, Enhanced Confix Stripping, Sastrawi, and Tala.

Researchers have explored various stemming algorithms to address the complexity of the Ambonese language. Stemming, defined as the process of reducing inflection or derivation to its base form, akin to transforming "nyapu" to "sapu," is commonly used for preprocessing in text-based applications (Sinaga & Nainggolan, 2023). Stemming algorithms are typically categorized into two types: statistical-based and rule-based. Statistical-based stemmers employ unsupervised algorithms leveraging training data to build models for stemming, while rule-based stemmers use predefined rules for stemming processes. Stemming can be applied to reduce search engine development and plagiarism (Tuhpatussania et al., 2022). However, challenges such as over-stemming and under-stemming commonly occur during the stemming process. It is essential to note that each language has unique characteristics and structures, particularly regarding affix structures, necessitating adjustments in stemming methods to fit the language's characteristics, and stemming methods differ across languages (Sovia et al., 2022).

Each language has different stemming algorithms, distinguishing them from those used in other languages. The Nazief-Adriani method, named after its authors, is exclusively designed and widely used for Indonesian stemming processes (Jumadi et al., 2021). This method uses rules and heuristics to stem by removing prefixes and suffixes from words. The Tala algorithm, created by Atmaja and Purwarianti, is another stemming method tailored for Indonesians. The Tala algorithm employs the Porter algorithm and follows rule-based operation principles (Pamungkas et al., 2023). This algorithm generates word stems using dictionary-based techniques and linguistic rules. The Sastrawi algorithm is also commonly used for stemming in Indonesian. Sastrawi is a stemmer algorithm that addresses the challenge of converting words into simpler forms (Rosid et al., 2020). This algorithm employs a combination of dictionary lookup and rule-based stemming to determine the base form of words. The Enhanced Confix Stripping algorithm is an improved version of the original Confix Stripping with modifications in word truncation (Jauhari et al., 2020). This algorithm is designed explicitly for Indonesians and utilizes a set of rules and dictionaries for stemming, focusing on affix removal and dictionary lookup.

Research findings on Javanese (Melia et al., 2023), Madurese (Lindrawati et al., 2023b), Balinese (Nata, 2023), and Minangkabau (Sovia et al., 2022) languages highlight both the effectiveness and limitations of existing stemming algorithms in their respective linguistic contexts. These studies provide insights into the suitability of these algorithms and areas where they may require refinement. However, when applied to Ambonese, challenges arise, indicating the need for further research and adaptation to enhance their effectiveness in this specific linguistic domain.

This literature review aims to conduct a comprehensive examination and comparative assessment of various stemming algorithms for text processing in Ambonese, aiming to address knowledge gaps and foster advancements in technology tailored to the linguistic characteristics of Ambonese. By integrating existing research, this review seeks to delineate these algorithms' strengths, weaknesses, and prospective applications in Ambonese text processing. This review aims to guide future research efforts by carefully analyzing and synthesizing relevant literature. The outcomes of this study can serve as a reference for selecting suitable stemming methods for indigenous languages, mainly Ambonese Malay, contributing to the development of linguistic research and application in Indonesia.

METHODS

Systematic Literature Review (SLR) endeavors to identify critical, relevant studies, gather necessary data, and then assess and combine findings to understand the research subject better (Saputra et al.,

2024). Systematic Literature Review (SLR) (Karuniawati et al., 2023) is outlined below regardless of a specific subject, academic field, or theoretical perspective (Figure 1).

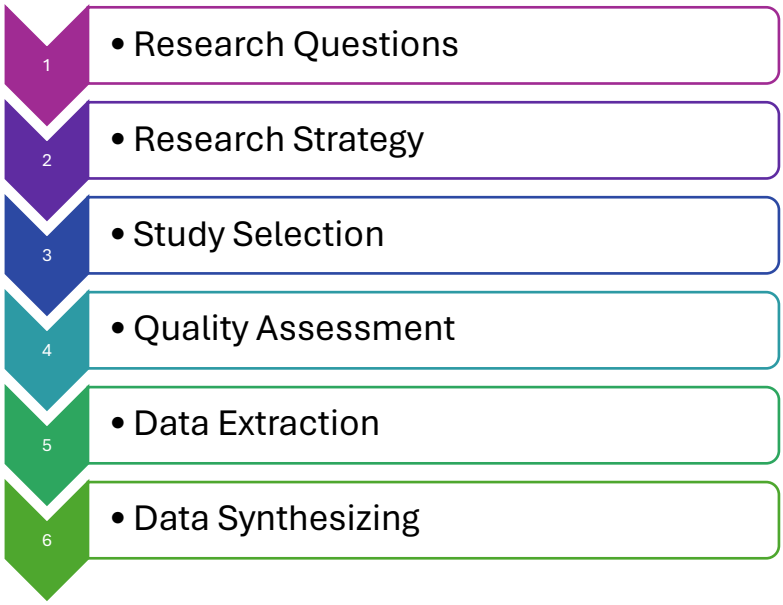

Figure 1. Research methods

Research Questions

Developing a series of Research Questions (RQs) is crucial when utilizing the systematic literature review (SLR) approach (Kusumah et al., 2024). The questions in Table 1 are crucial in shaping a clear, purposeful, and practical framework for research projects. This comprehensive approach aids in refining and focusing the research process, ultimately enhancing its effectiveness.

Table 1. Research questions

ID	Research Question
RQ1	What data-gathering approaches do researchers use in stemming-related studies?
RQ2	What methodologies are applied in the field of stemming research?
RQ3	What conclusions come from the analysis of stemming within the research context?

Research Strategy

A researcher thoroughly searches for scholarly papers across prominent databases such as ScienceDirect, IEEE Xplore, SpringerLink, Semantic Scholar, Google Scholar, and Elsevier. This search is conducted using several keywords, encompassing terms in both Indonesian and English, to ensure a comprehensive and inclusive retrieval of relevant literature:

- “Indonesian Stemming Algorithm”
- “Stemming Algorithm”
- “Nazief-Adriani Stemming Algorithm”
- “Sastrawi Algorithm”
- “Tala Algorithm”

Study Selection

In assessing manuscripts, it is crucial to establish criteria. The researcher utilizes two distinct types of criteria relevant to paper selection: inclusion criteria and exclusion criteria. The following are the specific inclusion criteria employed in the context of this study:

- Research paper is research conducted from 2019 to 2024.
- Research papers selected are written either in English or Indonesian.
- The main topic of the research study must be Indonesian stemming or Indonesian Regional Language Stemming.

Regarding the exclusion criteria, the researcher has already defined specific criteria for excluding papers from this study:

- Research that is not included in the inclusion criteria.
- The research does not clearly describe its flow or methodology.
- Research that fails to meet research objectives.

The criteria used in this research can be seen from the diagram of Figure 2 below:

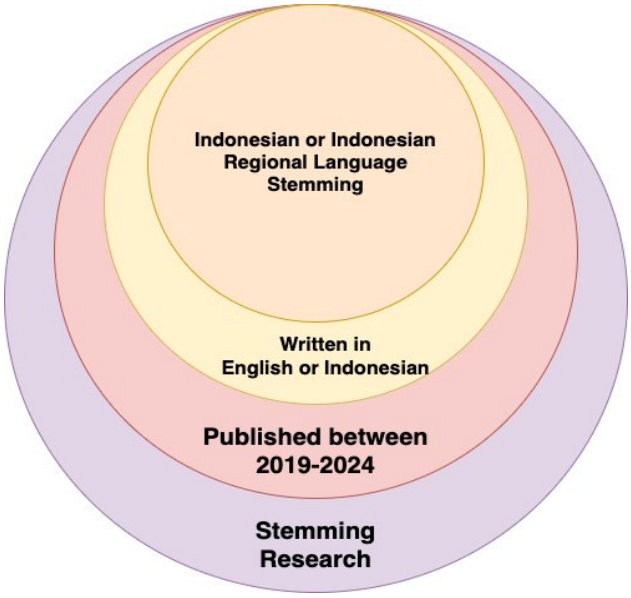

Figure 2. Research criteria

Quality Assessment

A meticulous quality assessment is imperative to achieve a comprehensive grasp of the overall quality of the study. This evaluation phase plays a crucial role in determining the relevance and suitability of the identified data for inclusion in the research. In the context of this study, the gathered data will undergo a rigorous evaluation process guided by a predefined set of criteria designed to gauge its quality. Employing these standards for quality assessment ensures a systematic and unbiased review, thereby enhancing the robustness and reliability of the study. Figure 3 shows the process of research paper selection.

Every paper selected for inclusion in this research must meet the following criteria:

- Was the research published between 2019 and 2024?
- Is the research written in Indonesian or English?
- Is the main research topic related to Indonesian or Indonesian regional language stemming?

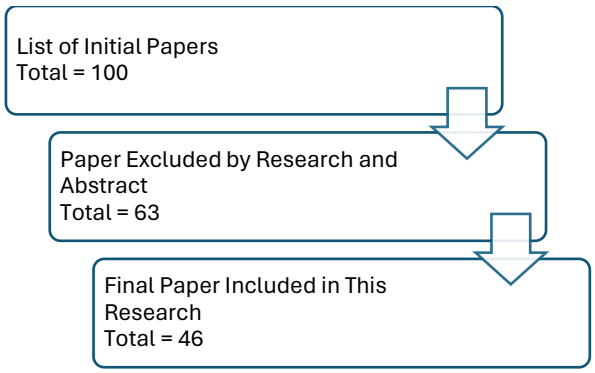

Figure 3. Research paper selection

Data Extraction

At this stage, the data extracted from the reviewed paper delves into crucial elements such as the publication year, the dataset utilized in the study under examination, the methodologies employed for data collection, the specific stemming approach adopted in the analyzed research, and the resulting implications of stemming on the study. Subsequently, all pertinent data was meticulously recorded in a spreadsheet document. This meticulous documentation serves as the cornerstone for comprehensively analyzing the gathered data. The systematic organization of this information facilitates an effective and structured investigation that adheres to established scholarly research standards.

Data Synthesizing

At this point in the research process, we have gathered 100 studies, carefully sifting through titles and abstracts for relevance. This initial screening narrowed down our selection to 63 papers. However, our scrutiny did not stop there. We meticulously handpicked 63 papers aligned with our research goals using stringent inclusion and exclusion criteria. If a paper met our inclusion criteria, it earned a spot in our literature review; conversely, those meeting exclusion criteria were omitted. This discerning process resulted in a final set of 34 papers, which underwent a detailed review and analysis. The gleaned data and primary findings from these papers underwent a thorough examination, and their synthesis is methodically presented in Table 2. This approach adheres to the scholarly standards of research, ensuring a comprehensive and well-structured exploration of the literature.

Table 2. Reviewed paper

Research	Research Year	Research Data	Stemming Method
(Yong et al., 2024)	2024	Data from web	Sastrawi
(Cahyaningrum et al., 2024)	2024	Data from web	Sastrawi
(Saifullah et al., 2024)	2023	Printed document	Sastrawi
(Fahreza et al., 2024)	2024	Data from web	Nazief & Adriani
(Cahyaningrum et al., 2024)	2024	Data from web	Nazief & Adriani
(Pamungkas et al., 2023)	2023	Data from web	Tala
(Nata, 2023)	2023	Corpus or dictionary	Tala
(Purwati et al., 2023)	2023	Corpus or dictionary	Sastrawi
(Bahtiar et al., 2023)	2023	Data from web	Sastrawi
(Xu et al., 2023)	2023	Data from web	Sastrawi

Research	Research Year	Research Data	Stemming Method
(Maylawati, Kumar, & Kasmin, 2023)	2023	Data from web	Sastrawi
(Lisangan et al., 2023)	2023	Printed document	Sastrawi
(Sinaga & Nainggolan, 2023)	2023	Printed document	Nazief & Adriani
(Pamungkas et al., 2023)	2023	Data from web	Nazief & Adriani
(Nata, 2023)	2023	Corpus or dictionary	Nazief & Adriani
(Maylawati, Kumar, & Kasmin, 2023)	2023	Corpus or dictionary	Nazief & Adriani
(Maylawati, Kumar, & Binti Kasmin, 2023)	2023	Printed document	Nazief & Adriani
(Lindrawati et al., 2023b)	2023	Corpus or dictionary	Nazief & Adriani
(Lindrawati et al., 2023a)	2023	Corpus or dictionary	Nazief & Adriani
(Chaidir, 2023)	2023	Data from web	Nazief & Adriani
(Aditya & Sumadi, 2023)	2023	Printed document	Nazief & Adriani
(Yaman et al., 2022)	2022	Printed document	Nazief & Adriani
(Tuhpatussania et al., 2022)	2022	Printed document	Nazief & Adriani
(Suzanti & Jauhari, 2022)	2022	Printed document	Nazief & Adriani
(Sovia et al., 2022)	2022	Corpus or dictionary	Nazief & Adriani
(Jaya Hidayat et al., 2022)	2022	Data from web	Nazief & Adriani
(Firman Sodiq et al., 2022)	2022	Data from web	Nazief & Adriani
(Amalia et al., 2022)	2022	Data from web	Nazief & Adriani
(Jaya Hidayat et al., 2022)	2022	Data from web	Sastrawi
(Tjut Adek et al., 2021)	2021	Data from web	Nazief & Adriani
(Rika Rosnelly et al., 2021)	2021	Printed document	Nazief & Adriani
(Prismana et al., 2021)	2021	Data from web	Nazief & Adriani
(Mustikasari et al., 2021)	2021	Corpus or dictionary	Nazief & Adriani
(Mahajan & Ingle, 2021)	2021	Data from web	Nazief & Adriani
(Jumadi et al., 2021)	2021	Corpus or dictionary	Nazief & Adriani
(Alfian et al., 2021)	2021	Corpus or dictionary	Nazief & Adriani
(Siswanto & Dani, 2021)	2021	Data from web	Sastrawi
(Rianto et al., 2020)	2020	Corpus or dictionary	Sastrawi
(Purbolaksono et al., 2020)	2020	Printed document	Sastrawi
(Fahmi et al., 2020)	2020	Data from web	Sastrawi
(Yunmar et al., 2020)	2020	Printed document	Nazief & Adriani
(Wibawa et al., 2020)	2020	Corpus or dictionary	Nazief & Adriani
(Purbolaksono et al., 2020)	2020	Printed document	Sastrawi
(Fahmi et al., 2020)	2020	Data from web	Sastrawi
(Soyusiawaty et al., 2020)	2020	Corpus or dictionary	Nazief & Adriani
(Simanjuntak et al., 2020)	2020	Printed document	Nazief & Adriani
(Yudhana et al., 2019)	2019	Corpus or dictionary	Nazief & Adriani

RESULTS AND DISCUSSION

Research Year

The collected papers will be grouped and classified according to their publication year as part of the analysis. This chronological approach will allow for a more nuanced assessment of the evolution of research trends and advances in text processing within the context of the Ambon language or other Indonesian Regional Languages. Each year's corpus of literature will be meticulously examined,

emphasizing identifying significant findings, methodology, and implications for stemming algorithms in the Ambon language.

By deconstructing the literature annually, this study hopes to shed light on the evolution of research in text processing for the Ambon language, providing significant insights for future research initiatives. Through this systematic exploration, we hope to improve text processing technologies tailored to the Ambon language's unique linguistic characteristics, fostering a better understanding and application of these algorithms in linguistic research and practice, as seen in Figure 4.

Figure 4. Research paper published year

In 2019, the research landscape witnessed the publication of 1 paper, marking the initial foray into exploring text-processing algorithms tailored to the linguistic nuances of the Ambon language. This modest beginning laid the groundwork for subsequent advancements in the field.

2020 witnessed a notable increase in research output, with nine papers published. This uptick reflects a growing interest and investment in studying text processing techniques and their applicability to the Ambon language. A researcher began delving deeper into the intricacies of stemming algorithms, aiming to unlock their full potential in linguistic analysis and application.

In 2021, the momentum continued with the publication of 8 papers. Despite a slight decrease compared to the previous year, this steady output underscores ongoing efforts to explore and refine text processing methodologies for the Ambon language.

The year 2022 has steadily grown, with eight papers published. This surge in research output signifies a maturing field as researchers delve into more nuanced aspects of text processing and strive to address lingering challenges and limitations.

In 2023, the upward trajectory continued, with 17 papers published. This consistent output indicates sustained interest and investment in advancing text-processing technologies tailored to the linguistic characteristics of the Ambon language.

However, in 2024, the number of published papers saw a notable decline, with only three papers making their way to publication. While this decrease may seem concerning at first glance, it is essential to consider potential factors such as ongoing research projects, funding availability, and external circumstances that may have influenced research output during this period.

Data Collection Techniques

When embarking on a study, researcher rely on data to guide their investigation. Data collection involves obtaining and analyzing precise data from various sources to address research questions, uncover patterns, explore possibilities, and assess potential outcomes. Throughout this process, the

researcher must delineate the nature of the data, ascertain its origins, and elucidate the methodology employed.

Data collection methods are diverse, with different strategies employed depending on the context. Notably, the scientific, commercial, and governmental sectors heavily rely on effective data-gathering procedures to inform their respective endeavors. This comprehensive approach aligns with the rigorous standards upheld in scientific research, ensuring the validity and reliability of the findings, as seen in Figure 5.

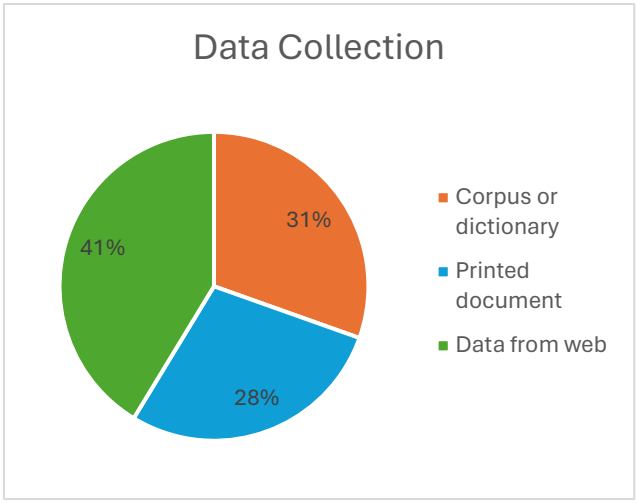

Figure 5. Research paper data collection

Much of the data collection effort is dedicated to compiling a corpus or dictionary, with 14 instances recorded. This method involves assembling a comprehensive collection of linguistic resources, such as texts, documents, and recordings, to serve as a basis for analysis and study. The corpus or dictionary is a rich source of authentic language data, allowing researchers to explore patterns, trends, and linguistic features within the Ambon language.

Printed documents constitute a substantial portion of the data collected, totaling 13 instances. This method involves sourcing information from published materials related to text processing, linguistics, and the Ambon language, such as books, journals, and articles. Printed documents provide valuable insights into existing research, theories, and methodologies as a foundation for building upon previous knowledge and expanding the scope of inquiry.

Additionally, data was sourced from web accounts for 19 instances of data collection. This method involves accessing and extracting information from online sources like websites, databases, and digital repositories. The web offers various resources, ranging from scholarly articles and research papers to linguistic databases and community forums. Leveraging web data enables researchers to access up-to-date information, diverse perspectives, and real-world examples relevant to their study of text processing in the Ambon language.

Researchers employ various data-gathering approaches in stemming-related studies. These include compiling corpora or dictionaries, which involve assembling linguistic resources to analyze language patterns. Printed documents, such as books and articles, offer insights into existing research and methodologies. Additionally, the researcher gathers data from the web, accessing diverse online sources for up-to-date information and real-world examples relevant to their study.

Stemming Methods

From the analysis, several stemming algorithms have been detected. These include the Nazief-Adriani algorithms, Sastrawi, and Tala, each offering distinct techniques for handling the intricacies of stemming in the Indonesian language. The selection of an appropriate algorithm is paramount in ensuring the accuracy and reliability of the stemming process within the analytical framework, as seen in Figure 6.

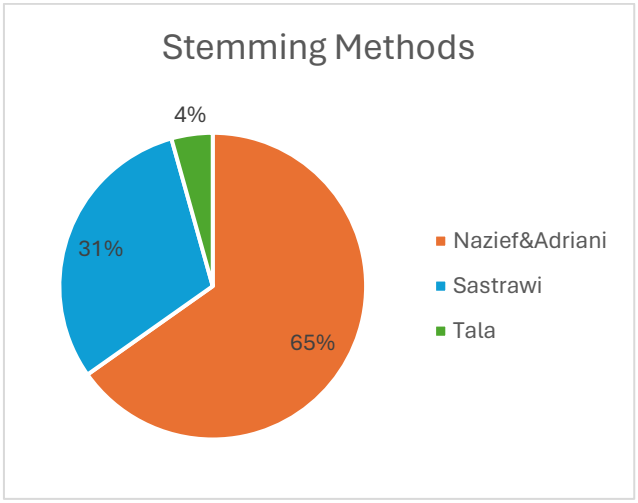

Figure 6. Research paper stemming methods

Nazief & Adriani algorithm emerges as the most frequently employed stemming method, appearing in 30 instances across the selected papers. This algorithm, named after its developers, offers a robust framework for stemming by employing rules and heuristics to extract root words from words with affixes, contributing to its widespread adoption and applicability.

Enhanced Confix Stripping is another prominent stemming method utilized in the selected papers, with 14 instances recorded. This enhanced version of the Confix Stripping algorithm focuses on affix stripping and dictionary lookup to perform stemming, offering an effective solution for handling the complexities of Indonesian language stemming.

Sastrawi is also featured in the papers, albeit to a lesser extent, with four instances recorded. This stemming algorithm combines dictionary lookup and rules-based stemming to derive the base form of words, providing the researcher with an additional linguistic analysis and processing tool.

Lastly, with two instances recorded, the Tala algorithm makes a limited appearance in the selected papers. Developed by Atmaja and Purwarianti, Tala utilizes a dictionary-based approach combined with linguistic rules to generate stems, offering a unique perspective on addressing the challenges of stemming in the Indonesian language.

Stemming Usage

The data shows that stemming is predominantly utilized in sentiment analysis, text processing, and document classification. These research areas often require preprocessing of textual data to extract meaningful features and improve the efficiency and effectiveness of subsequent analysis or classification tasks.

Additionally, stemming is employed in translation tasks, where the normalization of words through stemming aids in improving the accuracy and fluency of translated text. Furthermore, stemming finds

application in information retrieval systems, which facilitates matching user queries with relevant documents by reducing variations of words to their common root forms.

Moreover, the data suggests a significant focus on developing stemming algorithms. This indicates ongoing efforts to enhance the performance and adaptability of stemming techniques to various languages and linguistic contexts, including the Indonesian language. Stemming across diverse research areas underscores its importance as a fundamental preprocessing step in text analysis and information-processing tasks.

CONCLUSION

The research on stemming in the Indonesian language context, encompassing various stemming methods such as the Nazief & Adriani algorithm, Sastrawi, and Tala, reflects its pivotal role across diverse research domains, including sentiment analysis, text processing, document classification, translation, and information retrieval. Analysis of the data, comprising 30 instances of Nazief & Adriani, 14 of Sastrawi, and 2 of Tala implementations, indicates a significant utilization of stemming techniques, particularly in sentiment analysis, text processing, and document classification tasks, underscoring its efficacy in enhancing efficiency and accuracy. Moreover, stemming in translation tasks and information retrieval systems underscores its versatility and cross-domain applicability. The observation of 14 instances sourced from corpora or dictionaries, 13 from printed documents, and 19 from web data signify the diverse sources utilized in research. With a publication trend showing an increase in papers from 2019 to 2023 and a subsequent slight decline in 2024, the commitment to advancing stemming techniques is evident. This research aims to serve as a reference for selecting suitable stemming methods for indigenous languages, mainly Ambonese Malay, contributing to developing linguistic research and application in Indonesia.

REFERENCES

Aditya, C. S. K., & Sumadi, F. D. S. (2023). Combination of term weighting with class distribution and centroid-based approach for document classification. *Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control J*, *8*(4), 781–788. https://doi.org/10.22219/kinetik.v8i4.1793

Alfian, M., Barakbah, A. R., & Winarno, I. (2021). Indonesian online news extraction and clustering using evolving clustering. *JOIV:International Journal on Informatics Visualization*, *5*(3), 280. https://doi.org/10.30630/joiv.5.3.537

Amalia, A., Lidya, M. S., Andrian, A., Zamzami, E. M., & Hardi, S. M. (2022). OLCBot: Dissemination of interactive information related to Indonesia's omnibus law with implementing fuzzy string matching algorithm and sastrawi stemmer. *2022 6th International Conference on Electrical, Telecommunication and Computer Engineering (ELTICOM)*, 178–181. https://doi.org/10.1109/ELTICOM57747.2022.10037966

Bahtiar, S. A. H., Dewa, C. K., & Luthfi, A. (2023). Comparison of naïve Bayes and logistic regression in sentiment analysis on marketplace reviews using rating-based labeling. *Journal of Information Systems and Informatics*, *5*(3), 915–927. https://doi.org/10.51519/journalisi.v5i3.539

Cahyaningrum, L., Luthfiarta, A., & Rahayu, M. (2024). Sentiment analysis on the impact of mbkm on student organizations using supervised learning with smote to handle data imbalance. *Inform : Jurnal Ilmiah Bidang Teknologi Informasi Dan Komunikasi*, *9*(1). https://doi.org/10.25139/inform.v9i1.7484

Chaidir, I. (2023). Collaboration of Nazief & Adriani stemming algorithm with Postgresql queries parsing method to search for new study program names. *CESS (Journal of Computer Engineering, System and Science)*, *8*(2), 483. https://doi.org/10.24114/cess.v8i2.48212

Fahmi, S., Purnamawati, L., Shidik, G. F., Muljono, M., & Fanani, A. Z. (2020). Sentiment analysis of student review in learning management system based on sastrawi stemmer and SVM-PSO. *2020 International Seminar on Application for Technology of Information and Communication (iSemantic)*, 643–648. https://doi.org/10.1109/iSemantic50169.2020.9234291

Fahreza, M. D. A., Luthfiarta, A., Rafid, M., & Indrawan, M. (2024). Analisis sentimen: pengaruh jam kerja terhadap kesehatan mental generasi z. *Journal of Applied Computer Science and Technology*, *5*(1), 16–25. https://doi.org/10.52158/jacost.v5i1.715

Firman Sodiq, S., Desena, W., & Wibowo, A. (2022). Penerapan algoritma stemming Nazief & Adriani pada proses klasterisasi berita berdasarkan tematik pada laman (web) direktorat jenderal ham menggunakan rapidminer. *Syntax : Jurnal Informatika*, *11*(02), 10–21. https://doi.org/10.35706/syji.v11i02.7192

Jauhari, A., Suzanti, I. O., Pramudita, Y. D., Husni, & Diantisari, N. P. W. (2020). Enhanced confix stripping stemmer and cosine similarity for search engines in the holy Qur'an translation. *2020 6th Information Technology International Seminar (ITIS)*, 207–212. https://doi.org/10.1109/ITIS50118.2020.9321041

Jaya Hidayat, T. H., Ruldeviyani, Y., Aditama, A. R., Madya, G. R., Nugraha, A. W., & Adisaputra, M. W. (2022). Sentiment analysis of Twitter data related to Rinca Island development using Doc2Vec and SVM and logistic regression as a classifier. *Procedia Computer Science*, *197*, 660–667. https://doi.org/10.1016/j.procs.2021.12.187

Jumadi, J., Maylawati, D. S., Pratiwi, L. D., & Ramdhani, M. A. (2021). Comparison of Nazief-Adriani and Paice-Husk algorithm for Indonesian text stemming process. *IOP Conference Series: Materials Science and Engineering*, *1098*(3), 032044. https://doi.org/10.1088/1757-899X/1098/3/032044

Karuniawati, Y., Utami, E., & Yaqin, A. (2023). A Systematic Literature Review of Stemming in Non-Formal Indonesian Language. *IJISRT*, *8*(1).

Kusumah, P. A. D., Kusrini Kusrini, & Kusnawi Kusnawi. (2024). *Optimizing data security: A literature review on implementing Beaufort Cipher for Vigenère Affine Cipher*. https://doi.org/10.5281/ZENODO.10685974

Lindrawati, E., Utami, E., & Yaqin, A. (2023a). ANoM STEMMER: Nazief & Andriani modification for Madurese stemming. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, *7*(6), 1341–1347. https://doi.org/10.29207/resti.v7i6.5086

Lindrawati, E., Utami, E., & Yaqin, A. (2023b). Comparison of modified Nazief & Adriani and modified enhanced confix stripping algorithms for Madurese Language Stemming. *INTENSIF: Jurnal Ilmiah Penelitian dan Penerapan Teknologi Sistem Informasi*, *7*(2), 276–289. https://doi.org/10.29407/intensif.v7i2.20103

Lisangan, E. A., Midyanti, D. M., Mukmin, C., & Tungadi, A. L. (2023). The faculty of information technology at Atma Jaya University of Makassar uses the automatic classification system for academic performance evaluation. *JUITA : Jurnal Informatika*, *11*(1), 1. https://doi.org/10.30595/juita.v11i1.14116

Mahajan, S. D., & Ingle, D. D. R. (2021). News classification using machine learning. *International Journal of Innovative Science and Research Technology (IJISRT)*, *6*(5).

Maylawati, D. S., Kumar, Y. J., & Binti Kasmin, F. (2023). Feature-based approach and sequential pattern mining to enhance the quality of Indonesian automatic text summarization. *Indonesian Journal of Electrical Engineering and Computer Science*, *30*(3), 1795. https://doi.org/10.11591/ijeecs.v30.i3.pp1795-1804

Maylawati, D. S., Kumar, Y. J., & Kasmin, F. B. (2023). *Combination of graph-based approach and sequential pattern mining for extractive text summarization with Indonesian language*. *9*(2).

Melia, S. I., Sholihah, J., Nisak, D., Juniaristha, I. S., & Ni'mah, A. T. (2023). The ngoko Javanese stemmer uses the Enhanced Confix stripping stemmer method. *Rekayasa*, *16*(1), 107–112. https://doi.org/10.21107/rekayasa.v16i1.19308

Meturan, T., Laraswati Laraswati, & Triani, L. N. (2023). Bahasa Ambon dan bahasa Indonesia: Analisis fonologi. *Sintaksis : Publikasi Para Ahli Bahasa Dan Sastra Inggris*, *1*(5), 54–64. https://doi.org/10.61132/sintaksis.v1i5.261

Mustikasari, D., Widaningrum, I., Arifin, R., & Putri, W. H. E. (2021). *Comparison of the effectiveness of stemming algorithms in Indonesian documents:* 2nd Borobudur International Symposium on Science and Technology (BIS-STE 2020), Magelang, Indonesia. https://doi.org/10.2991/aer.k.210810.025

Nata, G. N. M. (2023). Pengembangan algoritma stemmer bilingual Bali-Indonesia dengan rule-base. *Seminar Nasional Corinsindo*, 278–283.

Pamungkas, N., Udayanti, E. D., Indriyono, B. V., Mahmud, W., Mintorini, E., Wahyu Dorroty, A. N., & Quamila Putri, S. (2023). Comparison of stemming test results of tala algorithms with Nazief Adriani

in abstract documents and national news. *Inform : Jurnal Ilmiah Bidang Teknologi Informasi Dan Komunikasi*, *8*(1), 33–41. https://doi.org/10.25139/inform.v8i1.5569

Pesiwarissa, L. F. (2023). Cigulu-cigulu (teka-teki) masyarakat tutur bahasa melayu Ambon (kajian etnosemantik: suatu pendekatan awal). *Prosiding Konferensi Linguistik Tahunan Atma Jaya (KOLITA)*, *21*(21), 208–214. https://doi.org/10.25170/kolita.21.4851

Prismana, I., Prehanto, D., Dermawan, D., Herlingga, A., & Wibawa, S. (2021). Nazief & Adriani Stemming Algorithm With Cosine Similarity Method For Integrated Telegram Chatbots With Service. *IOP Conference Series: Materials Science and Engineering*, *1125*(1), 012039. https://doi.org/10.1088/1757-899X/1125/1/012039

Purbolaksono, M. D., Reskyadita, F. D., Adiwijaya, -, Suryani, A. A., & Huda, A. F. (2020). Indonesian text classification using back propagation and sastrawi stemming analysis with information gain for selection feature. *International Journal on Advanced Science, Engineering and Information Technology*, *10*(1), 234–238. https://doi.org/10.18517/ijaseit.10.1.8858

Purwati, Y., Utomo, F. S., Trinarsih, N., & Hidayatulloh, H. (2023). Feature selection technique to improve the instances classification framework performance for Quran ontology. *JOIV : International Journal on Informatics Visualization*, *7*(2), 615. https://doi.org/10.30630/joiv.7.2.1195

Rianto, R., Mutiara, A. B., Wibowo, E. P., & Santosa, P. I. (2020). *Improving the accuracy of text classification using stemming method, a case of informal indonesian conversation*. https://doi.org/10.21203/rs.3.rs-41431/v1

Rika Rosnelly, Dedy Hartama, Muhammad Sadikin, & Cindy Paramitha Lubis. (2021). The similarity of essay examination results using preprocessing text mining with cosine similarity and Nazief-Adriani algorithms. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, *12*(3), 1415–1422. https://doi.org/10.17762/turcomat.v12i3.938

Rosid, M. A., Fitrani, A. S., Astutik, I. R. I., Mulloh, N. I., & Gozali, H. A. (2020). Improving Text preprocessing for student complaint document classification using Sastrawi. *IOP Conference Series: Materials Science and Engineering*, *874*(1), 012017. https://doi.org/10.1088/1757-899X/874/1/012017

Saifullah, S., Drężewski, R., Dwiyanto, F. A., Aribowo, A. S., Fauziah, Y., & Cahyana, N. H. (2024). Automated text annotation using a semi-supervised approach with meta vectorizer and machine learning algorithms for hate speech detection. *Applied Sciences*, *14*(3), 1078. https://doi.org/10.3390/app14031078

Saputra, W. A. M., Utami, E., & Yaqin, A. (2024). Unlocking insights: A literature review on enhanced confix stripping and Nazief & Adriani algorithm modifications for Makassar language text stemming. *International Journal of Innovative Science and Research Technology (IJISRT)*, 603–610. https://doi.org/10.38124/ijisrt/IJISRT24MAR437

Simanjuntak, M. S., Panjaitan, J., & Syahputra, S. A. (2020). Using preprocessing text mining with Nazief-Adriani algorithms similarity of essay final exam semester. *Institute of Computer Science*, *4*(36).

Sinaga, A., & Nainggolan, S. P. (2023). Analisis perbandingan akurasi dan waktu proses algoritma Stemming Arifin-Setiono dan Nazief-Adriani pada dokumen teks bahasa Indonesia. *Sebatik*, *27*(1), 63–69. https://doi.org/10.46984/sebatik.v27i1.2072

Siswanto, B., & Dani, Y. (2021). Sentiment analysis about oximeter as covid-19 detection tools on Twitter using sastrawi library. *2021 8th International Conference on Information Technology, Computer and Electrical Engineering (ICITACEE)*, 161–164. https://doi.org/10.1109/ICITACEE53184.2021.9617216

Sovia, R., Defit, S., & Yuhandri. (2022). Development of the Minangkabau local language translation machine based on stemming. *2022 International Symposium on Information Technology and Digital Innovation (ISITDI)*, 195–198. https://doi.org/10.1109/ISITDI55734.2022.9944457

Soyusiawaty, D., Jones, A. H. S., & Lestariw, N. L. (2020). The stemming application on affixed Javanese words by using the Nazief and Adriani algorithm. *IOP Conference Series: Materials Science and Engineering*, *771*(1), 012026. https://doi.org/10.1088/1757-899X/771/1/012026

Suzanti, I. O., & Jauhari, A. (2022). Comparison of stemming and similarity algorithms in Indonesian translated Al-Qur'an text search. *Jurnal Ilmiah Kursor*, *11*(2), 91. https://doi.org/10.21107/kursor.v11i2.280

Tjut Adek, R., Kesuma Dinata, R., & Ditha, A. (2021). Online newspaper clustering in Aceh using the agglomerative hierarchical clustering method. *International Journal of Engineering, Science, and Information Technology*, *2*(1), 70–75. https://doi.org/10.52088/ijesty.v2i1.206

Tuhpatussania, S., Utami, E., & Hartanto, A. D. (2022). Comparison of porter stemming algorithm and Nazief & Adriani's stemming algorithm in determining Indonesian language learning modules. *Jurnal Pilar Nusa Mandiri*, *18*(2), 203–210. https://doi.org/10.33480/pilar.v18i2.3940

Wibawa, A. P., Dwiyanto, F. A., Zaeni, I. A. E., Nurrohman, R. K., & Afandi, A. (2020). Stemming Javanese affix words using Nazief and Adriani modifications. *Jurnal Informatika*, *14*(1), 36. https://doi.org/10.26555/jifo.v14i1.a17106

Xu, A., Tiffany, T., Phanie, M. E., & Simarmata, A. (2023). Sentiment analysis on Twitter posts about the Russian and Ukrainian war with long short-term memory. *SinkrOn*, *8*(2), 789–797. https://doi.org/10.33395/sinkron.v8i2.12235

Yaman, A., Sartono, B., Indrawati, A., Kartika, Y. A., & Soleh, A. M. (2022). Automated multi-label classification on fertilizer-themed patent documents in Indonesia. *DESIDOC Journal of Library & Information Technology*, *42*(4), 218–226. https://doi.org/10.14429/djlit.42.4.17733

Yong, J. Z. H., Koh, J. Y., Liew, J. X., & Tan, C. W. (2024). Linguistic harmony in diversity: lemmatizing rojak Malay for global communication. *2024 3rd International Conference on Digital Transformation and Applications (ICDXA)*, 6–10. https://doi.org/10.1109/ICDXA61007.2024.10470819

Yudhana, A., Fadlil, A., & Rosidin, M. (2019). Indonesian words error detection system using Nazief Adriani stemmer algorithm. *International Journal of Advanced Computer Science and Applications*, *10*(12). https://doi.org/10.14569/IJACSA.2019.0101231

Yunmar, R. A., Setiawan, A., & Tantriawan, H. (2020). The combination of Yake and language processing for unsupervised term extraction ontology learning. *IOP Conference Series: Earth and Environmental Science*, *537*(1), 012023. https://doi.org/10.1088/1755-1315/537/1/012023