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

Vinnesa Patricia Carolina¹, Ema Utami², Ainul Yaqin³

¹²³Magister Informatika, Universitas AMIKOM Yogyakarta, Sleman, Indonesia

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.

This is an open-access article under the CC-BY-SA license.

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

![Image](image_url)

**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>
<thead>
<tr>
<th>ID</th>
<th>Research Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>RQ1</td>
<td>What data-gathering approaches do researchers use in stemming-related studies?</td>
</tr>
<tr>
<td>RQ2</td>
<td>What methodologies are applied in the field of stemming research?</td>
</tr>
<tr>
<td>RQ3</td>
<td>What conclusions come from the analysis of stemming within the research context?</td>
</tr>
</tbody>
</table>

**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:

![Research criteria diagram]

**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?
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>
<thead>
<tr>
<th>Table 2. Reviewed paper</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Yong et al., 2024)</td>
</tr>
<tr>
<td>(Cahyaningrum et al., 2024)</td>
</tr>
<tr>
<td>(Saifullah et al., 2024)</td>
</tr>
<tr>
<td>(Fahreza et al., 2024)</td>
</tr>
<tr>
<td>(Cahyaningrum et al., 2024)</td>
</tr>
<tr>
<td>(Pamungkas et al., 2023)</td>
</tr>
<tr>
<td>(Nata, 2023)</td>
</tr>
<tr>
<td>(Purwati et al., 2023)</td>
</tr>
<tr>
<td>(Bahtiar et al., 2023)</td>
</tr>
<tr>
<td>(Xu et al., 2023)</td>
</tr>
</tbody>
</table>
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](image)

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
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.

![Data Collection Chart]

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](image)

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**


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).


