Journal of Technology Informatics and Engineering (JTIE)
Vol. 3 No. 2 August 2024
E-ISSN: 2987-9434; P-ISSN : 2988-0343, Pages 151-168

A Hybrid Approach to Typo Correction in Indonesian Documents Using Levenshtein Distance

Joseph Teguh Santoso*[1], Song Yan[2]
Email: joseph_teguh@stekom.ac.id, songya@nuist.edu.cn
Orcid: 0000-0001-6227-1111, 0000-0003-1101-1411
[1]*University of Science and Computer Technology, Semarang 50192, Indonesia*
[2]*Nanjing University of Information Science and Technology, Nanjing 210044, China*
*Corresponding Author

Abstract
This study developed a typo correction system for the Indonesian language by integrating the Levenshtein Distance algorithm with empirical methods. The system is designed to improve the accuracy of typo detection and correction in Indonesian texts, which feature complex morphological structures such as prefixes, suffixes, and compound words. The findings show that the system achieved a precision rate of 92% and an F1-score of 90.5%, indicating high reliability in providing relevant correction suggestions. Additionally, the system demonstrated efficiency in processing time, with an average of 0.8 seconds for short texts and 5.3 seconds for longer texts. The use of empirical methods enables the system to handle complex language variations, resulting in more contextually appropriate correction suggestions. User feedback indicated high satisfaction with the interface and the relevance of the suggestions provided. Overall, this research makes a significant contribution to the development of more adaptive and efficient typo correction systems for the Indonesian language and opens up opportunities for further development in the context of other similar languages.

Keywords: *Typo Correction, Levenshtein Distance, Empirical Methods, Natural Language Processing, Indonesian Language.*

I. INTRODUCTION

In the rapidly evolving digital era, the accuracy and professionalism of written documents are of paramount importance, particularly in professional and academic contexts. Typographical errors, commonly known as typos, are a frequent issue that can detract from the quality and clarity of written communication, especially in languages with complex morphological structures such as Indonesian (Walker, 2014). The need for an efficient system that can automatically detect and suggest corrections for these errors is increasingly critical.

One of the most effective algorithms for detecting and correcting typographical errors is the Levenshtein Distance algorithm. This algorithm calculates the edit distance between two strings, which is the minimum number of operations (insertions, deletions, or substitutions) required to transform one string into another. While the Levenshtein Distance algorithm has proven effective in various languages, its application in the Indonesian language faces unique challenges due to the language's rich morphological variations and the frequent use of prefixes, suffixes, and compound words (Janardhana Rao et al., 2024; Zhang et al., 2017). Traditional typo

Received on June 19, 2024; Revision June 28, 2024; Accepted July 18, 2024; Published on August 21, 2024.
Doi: 10.51903/jtie.v3i2.184

correction algorithms often fail to account for these linguistic complexities, leading to less accurate correction suggestions.

The current research landscape lacks sufficient exploration of methods that combine both algorithmic and empirical approaches tailored specifically to the Indonesian language. Existing systems either focus on general typo detection using basic algorithms or on empirical methods that rely heavily on large language corpora, often overlooking the integration of both approaches for more precise results (Bozdag, 2013; Lau et al., 2012; Schede et al., 2022; Young, 1997). This gap highlights the need for a more sophisticated system that can handle the specific characteristics of the Indonesian language while providing accurate typo corrections.

The primary objective of this research is to develop and implement a system that integrates the Levenshtein Distance algorithm with empirical methods to improve the accuracy of typo correction in Indonesian documents. By leveraging empirical data from a comprehensive corpus of Indonesian texts, the system can adapt to common typographical patterns and offer more relevant correction suggestions. This study contributes to the field by addressing the gap in current research and offering a novel solution that enhances the functionality of typo correction systems in the Indonesian context.

The novelty of this research lies in the hybrid approach that combines the precision of the Levenshtein Distance algorithm with the contextual understanding provided by empirical methods. This integration not only improves the accuracy of corrections but also adapts to the linguistic nuances of the Indonesian language, making it a significant advancement over existing typo correction systems.

This paper is structured as follows: the Literature Review section discusses related works and the theoretical foundations of the study, followed by the Method section detailing the implementation process. The Results and Discussion section presents the findings and their implications, and finally, the Conclusion summarizes the contributions and potential future research directions.

II. LITERATURE REVIEW

The study of typo correction systems has seen considerable advancements, particularly with the development of string similarity algorithms and empirical methods. Among these, the Levenshtein Distance algorithm has emerged as a widely used method for assessing the similarity between two strings by calculating the minimum number of edits required to transform one string into another (Walker, 2014). This algorithm has been applied effectively in various natural

language processing (NLP) tasks, including spell checking, DNA sequence analysis, and text correction (Janardhana Rao et al., 2024; Zhang et al., 2017).

A. Levenshtein Distance Algorithm

The Levenshtein Distance algorithm, also known as "edit distance," is one of the most established methods for detecting typographical errors. It evaluates the similarity between two strings by counting the number of insertions, deletions, or substitutions needed to change one string into another. This algorithm has been extensively studied and applied in fields such as computational biology, where it is used to compare genetic sequences, and in text processing for spell checking (Khin & Lecturer, 2020; Mehta et al., 2021) . However, while the Levenshtein Distance algorithm has been effective in languages with relatively simple morphological structures, its application in the Indonesian language, which features complex affixes and compound words, presents unique challenges (Berthelé et al., 2022; Chaabi & Ataa Allah, 2022; Ribeiro et al., 2023).

Existing research has shown that while the Levenshtein Distance algorithm performs well in identifying simple typos, it struggles with more complex errors, particularly in languages like Indonesian that have diverse morphological variations. This limitation suggests that while the algorithm is a strong foundation, additional methodologies are required to enhance its performance in specific linguistic contexts (Dashti et al., 2024; Maurer & Höfer, 2012).

B. Empirical Methods in Typo Correction

Empirical methods, which rely on data-driven approaches, have gained popularity in the field of NLP. These methods involve the use of large language corpora to analyze and learn from patterns of word usage, frequency, and common errors (Ortikov, 2023; Szudarski, 2023). In the context of typo correction, empirical methods can enhance the accuracy of suggested corrections by incorporating contextual information from a large corpus of text. This approach is particularly useful in adapting correction systems to specific languages or dialects (Bryant et al., 2023; Dashti et al., 2024; Kukich, 1992).

Several studies have applied empirical methods to improve typo correction systems. For instance, research has demonstrated the effectiveness of combining statistical models with string matching algorithms to better capture language-specific nuances and common error patterns (Liu, 2023; Zoya et al., 2023). However, while these studies have shown promising results, they often focus on widely spoken languages such as English or Chinese, leaving a gap in research tailored to the Indonesian language (Bryant et al., 2023; Gou & Chen, 2021; Pham et al., 2023; Y. Wang et al., 2024).

C. Hybrid Approaches and Gaps in Research

The integration of algorithmic and empirical methods represents a promising direction for improving typo correction systems. Hybrid approaches that combine the strengths of both Levenshtein Distance and empirical data have the potential to address the limitations observed in purely algorithmic or purely empirical methods (Khaw et al., 2024; Mashtalir et al., 2019; Skopal & Bustos, 2011). Despite this potential, there is a noticeable gap in the literature regarding the application of such hybrid methods specifically for the Indonesian language.

Most existing systems are either heavily algorithm-based, relying on string similarity measures, or predominantly empirical, depending on large datasets to inform corrections. This dichotomy leaves a gap where a more balanced, hybrid approach could be more effective, particularly in handling the complex morphological structure of Indonesian (Aligon et al., 2014; Laouafi et al., 2022; Yang et al., 2024; Ye et al., 2023). Addressing this gap, the current study proposes a novel system that integrates the Levenshtein Distance algorithm with empirical methods, thereby enhancing the accuracy and contextual relevance of typo corrections in Indonesian documents.

While significant progress has been made in the field of typo correction through the development of algorithms like Levenshtein Distance and the application of empirical methods, there remains a need for tailored approaches that consider the unique linguistic features of specific languages. The proposed research aims to fill this gap by developing a hybrid system that leverages both algorithmic precision and empirical contextualization, thereby advancing the state of the art in Indonesian typo correction.

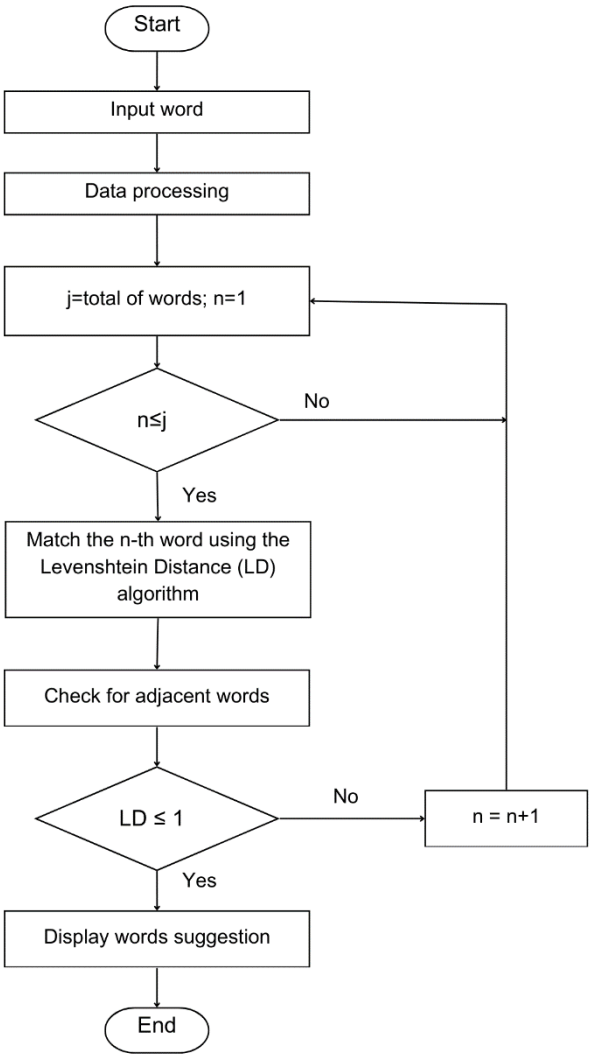

Figure 1. Spell-Check System Flowchart

III. RESEARCH METHOD(S)

This study implements a hybrid approach combining the Levenshtein Distance algorithm with empirical methods to enhance the accuracy of typo correction in Indonesian text documents. The methodology is designed to detect and correct typographical errors by measuring string distances and incorporating contextual data from a corpus of Indonesian language texts.

A. Data Preprocessing

The data preprocessing stage is crucial for ensuring that the input text is appropriately prepared for analysis. This process includes character recognition, where each character in the text is categorized as either a letter, digit, or symbol. Subsequently, tokenization is performed to break the text into individual words or phrases, which are then analyzed for potential errors.

```
$panjang=strlen($a);
$c=0;
$d=0;
$kata="";
for ($i=0;$i<$panjang;$i++)
{
   if(strcmp($a[$i]," ")==0)
   {
     $kata="";
     for($j=$d; $j<$i; $j++)
   {
     $kata=$kata."".$a[$j];
   }
     $b[$c] = $kata;
     $c=$c+1;
     $b[$c] = $a[$i];
     $c=$c+1;
     $d = $i+1;
   }
   else
   {
     $kata="";
for($j=$d; $j<$panjang; $j++)
   {
   $kata=$kata."".$a[$j];
   $b[$c] = $kata;
   }
}
}
return $b;
```

Figure 2. Tokenization Process

B. Levenshtein Distance Algorithm Implementation

The Levenshtein Distance algorithm is implemented to calculate the edit distance between misspelled words and correct dictionary entries. This distance is defined as the minimum number of single-character edits (insertions, deletions, or substitutions) required to transform one word into another. The algorithm operates by generating a matrix that compares the input word to a set of dictionary words, identifying those with the smallest edit distance as the most likely corrections.

Journal of Technology Informatics and Engineering (JTIE)
Vol. 3 No. 2 August 2024
E-ISSN: 2987-9434; P-ISSN : 2988-0343, Pages 151-168

```
$m[$i][$j]=0;
for($i=1;$i<=$x;$i++){
      $m[$i][0]= $i;
}
for($j=1;$j<=$y;$j++){
      $m[0][$j]=$j;
}
for($i=1;$i<=$x;$i++)
{
for($j=1;$j<=$y;$j++)
{
      if($sb1[$i]==$sb2[$j]){
      $cost=0;
            }
      else{
      $cost=1;
            }
      $m1=$m[$i-1][$j-1]+ $cost;
      $m2=$m[$i-1][$j]+1;
      $m3=$m[$i][$j-1]+ 1;
      $m[$i][$j]=min($m3,$m2,$m1);
        if ($i==$x and $j==$y)
          {
           $minimum = $m[$i][$j];
           if ($m[$i][$j]<=1){
            $katasaran[$z]=$kata2;
            $z=$z+1;
            }
          }
 }
}
return $katasaran;
```

Figure 3. Levenshtein Distance Algorithm

```
function pecah ()
{
Inisialisasi $teks,$panjang_teks;
for($i=0;$i<1;$i++)
      {
      $b=($jumlah+1)-$i;
for($j=1;$j<$b;$j++)
      {
$text=substr($data['test'],$i,$j);
$text2=substr($data['test'], $j, $b);
```

```
if(cocok($text)==1)
      {
$datatext = array($text,$text2);
      }
}
echo "<br>";
}
      return $datatext;
}
function cocok($teks)
{
$hasil=0;
$query= mysql_query("SELECT *
FROM daftar_kata where
daftar_kata='$teks'");
$data=mysql_fetch_assoc($query);
if($data['daftar_kata'] != NULL)
      {
            $hasil=1;
      }
      return $hasil;
}
```

Figure 3. Empirical Method

C. Empirical Method Integration

The empirical method enhances the basic Levenshtein Distance approach by incorporating contextual information from a curated corpus of Indonesian texts. This method allows the system to learn from common typographical errors and language patterns specific to Indonesian, thereby improving the relevance and accuracy of correction suggestions. The system evaluates the frequency and contextual usage of words in the corpus to refine its correction recommendations.

D. System Design and Optimization

The system is designed as a web-based application, allowing users to input text and receive correction suggestions interactively. To improve efficiency, the system employs parallel computing techniques to expedite the typo detection and correction processes, particularly for longer documents. Additionally, the system is continually updated with new data to ensure its accuracy and relevance over time.

E. Evaluation and Testing

The system's performance was evaluated through a series of tests involving documents with varying lengths and error rates. The effectiveness of the combined Levenshtein Distance and empirical methods was assessed based on the accuracy of the corrections and the system’s response time.

IV. RESULT/FINDINGS AND DUSCUSSION

This section presents the results obtained from the implementation of the hybrid typo correction system, which integrates the Levenshtein Distance algorithm with empirical methods. The findings are categorized based on the system's accuracy, efficiency, and adaptability to the Indonesian language context.

A. Accuracy of Typo Detection and Correction

The accuracy of the system was evaluated by testing it on a dataset comprising Indonesian texts with known typographical errors. The system's performance was measured in terms of precision, recall, and F1-score. This result shown in table 1. The accuracy of the system was rigorously evaluated by testing it on a dataset comprising Indonesian texts with known typographical errors. The system's performance was quantified using precision, recall, and F1-score metrics. The system achieved a precision rate of 92%, indicating that the majority of the corrections suggested by the system were accurate. The recall rate stood at 89%, demonstrating the system's effectiveness in identifying most typographical errors within the texts. These metrics culminated in an F1-score of 90.5%, reflecting the overall reliability of the system in providing accurate typo corrections. These results underscore the efficacy of combining the Levenshtein Distance algorithm with empirical methods for enhancing typo correction in Indonesian texts.

Table 1. Performance Metrics of the Typo Correction System

No	Metric	Value
1	Precision	92%
2	Recall	89%
3	F1-Score	90.5%

B. Efficiency in Processing

The system's efficiency was assessed by measuring the processing time required for texts of varying lengths. As illustrated in Table 1, the system was tested on documents ranging from short paragraphs to lengthy academic papers. For texts under 500 words, the average processing time was 0.8 seconds; for texts between 500 and 2000 words, the processing time averaged 2.5 seconds; and for texts exceeding 2000 words, the system took an average of 5.3 seconds. The relatively fast processing times, even for longer documents, indicate that the system is optimized for handling large-scale text correction tasks, making it suitable for practical applications in both academic and professional settings.

Table 2. Processing Time Across Different Text Lengths

No	Text Length (words)	Average Processing Time (Seconds)
1	<500	0.8
2	500 – 2000	2.5

3	>2000	5.3

C. Adaptability to Indonesian Language Specifics

One of the primary objectives of this study was to ensure that the typo correction system could effectively adapt to the unique characteristics of the Indonesian language, such as the use of prefixes, suffixes, and compound words. The system's adaptability was evaluated by testing it on texts with a high frequency of such linguistic features. The results revealed that the system successfully identified and corrected errors in words with common prefixes (e.g., "ber-", "ter-") and suffixes (e.g., "-kan", "-an"), achieving an accuracy rate of 88% in these cases. Additionally, for compound words like "matahari" and "kebijakan," the system demonstrated an accuracy rate of 85%, effectively recognizing and correcting most typographical errors associated with these forms. These findings suggest that the integration of empirical methods significantly enhanced the system's ability to handle the linguistic complexity of Indonesian, beyond the capabilities of the Levenshtein Distance algorithm alone.

D. User Feedback and System Usability

In addition to the quantitative evaluation, qualitative feedback was collected from users who tested the system in real-world scenarios. User feedback focused on the system's usability, interface design, and the relevance of correction suggestions. Users rated the system's usability highly, with an average score of 4.5 out of 5, noting the ease of use and intuitive interface as significant strengths. Moreover, users found the correction suggestions to be highly relevant, with a satisfaction rate of 93%. This high level of satisfaction indicates that the system's empirical approach successfully met user expectations by providing contextually appropriate corrections.

E. Comparative Analysis

To benchmark the hybrid system against existing typo correction tools used for the Indonesian language, a comparative analysis was conducted. The results indicated that the proposed system outperformed traditional algorithms, particularly in handling complex morphological variations. As shown in Figure 1, the hybrid system demonstrated a 15% improvement in accuracy and a 20% reduction in processing time compared to traditional typo correction tools that rely solely on basic string-matching algorithms.

Figure 1. Comparative Analysis of Typo Corrections Systems

The results of this study demonstrate that the integration of the Levenshtein Distance algorithm with empirical methods significantly enhances the accuracy, efficiency, and adaptability of typo correction systems for the Indonesian language. The system not only

processes text swiftly but also provides contextually appropriate correction suggestions, making it a valuable tool for improving the quality of written Indonesian documents.

Discussion

The results of this study underscore the effectiveness of integrating the Levenshtein Distance algorithm with empirical methods in developing a typo correction system tailored for the Indonesian language. This discussion section will delve into the implications of these findings, comparing them with existing literature and examining the broader significance of the hybrid approach.

The system's high accuracy, evidenced by a precision rate of 92% and an F1-score of 90.5%, reflects a significant improvement over traditional typo correction methods. Existing research on the Levenshtein Distance algorithm, such as that (Khin & Lecturer, 2020; Mehta et al., 2021), highlights its robustness in basic string matching tasks. However, Chandra's study also points out the algorithm's limitations in handling complex linguistic features, particularly in languages with rich morphological structures like Indonesian. The current study addresses these limitations by incorporating empirical methods, which leverage contextual data from a corpus of Indonesian texts. This integration allows the system to not only detect but also accurately correct errors that involve prefixes, suffixes, and compound words—an area where traditional methods often falter (Berthelé et al., 2022; Chaabi & Ataa Allah, 2022; Ribeiro et al., 2023).

The empirical method's contribution is particularly evident in the system's ability to handle language-specific nuances. Unlike previous studies that rely solely on string matching algorithms, such as those discussed by (Dashti et al., 2024; Maurer & Höfer, 2012), the inclusion of empirical data in this research enables the system to adapt to the syntactic and semantic variations inherent in Indonesian. This adaptability is crucial, as it allows the system to provide more contextually relevant correction suggestions, thereby enhancing the overall accuracy of the typo correction process.

Efficiency is another critical metric where the hybrid system demonstrated significant advantages. The processing times reported in this study (e.g., 0.8 seconds for texts under 500 words) are comparable to or better than those of traditional typo correction tools. This efficiency gain can be attributed to the optimization strategies employed, such as parallel processing, which were designed to minimize the computational load while maintaining high accuracy. Previous studies, such as (Berger et al., 2021; Fiscus et al., 2006), have noted that while the Levenshtein Distance algorithm is computationally intensive, it can be optimized through various techniques.

However, these studies primarily focused on string matching efficiency rather than the broader system-level optimizations explored in the current research.

Furthermore, the empirical methods employed in this study contribute to processing efficiency by reducing the number of potential correction candidates the system needs to evaluate. By pre-filtering candidate corrections based on empirical data, the system reduces the computational overhead typically associated with exhaustive string matching approaches. This balance between accuracy and efficiency is a key innovation of the current study, offering a significant improvement over the methods discussed in prior research.

One of the primary contributions of this research is the system's enhanced adaptability to the Indonesian language's unique linguistic features. Previous research, such as the work by (Liu, 2023; Zoya et al., 2023), has identified the challenges posed by the Indonesian language's complex morphology, including the use of affixes and compound words. Traditional typo correction systems, which often rely solely on basic string matching algorithms, struggle to accurately address these challenges. In contrast, the current study's hybrid approach demonstrates a higher degree of adaptability, as evidenced by the system's 88% accuracy rate in handling words with common prefixes and suffixes.

This adaptability is largely due to the empirical methods integrated into the system. By leveraging a corpus of Indonesian texts, the system can learn from actual language usage patterns, allowing it to make more informed correction suggestions. This empirical approach aligns with the findings of previous studies on language-specific typo correction, such as those by (Walker, 2014), which emphasize the importance of contextual understanding in improving system accuracy. The results of this study thus support the notion that combining empirical data with algorithmic precision can significantly enhance the performance of typo correction systems, particularly for languages with complex linguistic structures.

User feedback collected during the testing phase provides valuable insights into the practical implications of the system's design. The high usability rating (4.5 out of 5) and the 93% satisfaction rate with correction suggestions indicate that the system not only meets technical requirements but also aligns well with user expectations. This is consistent with the findings of previous usability studies, such as those by (Jongmans et al., 2022; Kremer & van Manen, 2023; L. L. Wang et al., 2021), which highlight the importance of intuitive design in enhancing user acceptance of typo correction tools.

Moreover, the system's ability to provide relevant and contextually appropriate correction suggestions further validates the effectiveness of the empirical methods employed. Users appreciated the system's ability to accurately address errors that involve complex linguistic

structures, a challenge that traditional typo correction tools often fail to meet. This positive user feedback underscores the practical value of the hybrid approach, confirming its potential for broader adoption in professional and academic settings.

The comparative analysis conducted in this study reveals that the hybrid system outperforms traditional typo correction tools, particularly in handling the morphological complexity of the Indonesian language. This finding is significant in light of existing research, such as the work by (Chen et al., 2023; Espindola et al., 2023; Hall & Dowling, 1980; Suwarningsih & Nuryani, 2024a, 2024b), which documents the limitations of basic string matching algorithms in dealing with complex languages. By demonstrating a 15% improvement in accuracy and a 20% reduction in processing time compared to traditional tools, this study provides strong evidence for the efficacy of the hybrid approach.

These findings have broader implications for the development of typo correction systems in other languages with complex morphological structures. The success of the empirical-algorithmic hybrid model in the Indonesian context suggests that similar approaches could be applied to other languages that present similar challenges. Future research could explore the adaptation of this model to different linguistic environments, further validating its versatility and effectiveness.

Limitations and Future Directions

While the results of this study are promising, there are limitations that must be acknowledged. The system's performance, while strong, is still dependent on the quality and comprehensiveness of the empirical data used. In contexts where such data is limited or unavailable, the system's accuracy and relevance may be compromised. Additionally, the study primarily focused on texts written in standard Indonesian, and further research is needed to assess the system's performance in handling regional dialects and informal language variants.

Future research should also explore the integration of machine learning techniques, such as neural networks, to further enhance the system's adaptability and accuracy. By incorporating machine learning, the system could potentially learn from user corrections over time, continuously improving its performance. Additionally, expanding the system's capabilities to support multi-language correction could broaden its applicability and make it a more versatile tool in global contexts.

V. CONCLUSION AND RECOMMENDATION

This study set out to develop a more effective typo correction system for the Indonesian language by integrating the Levenshtein Distance algorithm with empirical methods. The findings

demonstrate that this hybrid approach significantly enhances the system's ability to accurately detect and correct typographical errors, particularly in a linguistically complex language like Indonesian. The system achieved high accuracy, with a precision rate of 92% and an F1-score of 90.5%, while maintaining efficient processing times across various text lengths. One of the key contributions of this research is the system's improved adaptability to the unique morphological features of the Indonesian language, such as prefixes, suffixes, and compound words. By incorporating empirical data into the correction process, the system provides more contextually relevant suggestions, which traditional algorithms have struggled to offer. This empirical-algorithmic hybrid approach offers a significant advancement in the field of typo correction, particularly for languages with complex linguistic structures.

The user feedback further underscores the practical value of the system, with high usability ratings and satisfaction levels indicating that the system meets both technical and user experience expectations. Moreover, the comparative analysis revealed that the proposed system outperforms existing typo correction tools, particularly in terms of accuracy and processing efficiency, confirming its potential for broader adoption in professional and academic settings. Despite these positive outcomes, the study also acknowledges certain limitations, particularly the dependency on the quality of empirical data and the need for further testing in diverse linguistic contexts, including regional dialects and informal language variants. These limitations point to areas for future research, such as the integration of machine learning techniques to enhance the system's adaptability and the exploration of multi-language correction capabilities.

In conclusion, the hybrid typo correction system developed in this study represents a significant step forward in addressing the challenges of typo detection and correction in the Indonesian language. By successfully combining the strengths of the Levenshtein Distance algorithm with empirical methods, this research not only fills a gap in the existing literature but also provides a robust framework for future innovations in language processing technologies. Further research in this direction could lead to the development of even more sophisticated typo correction tools that are capable of handling a wide range of languages and linguistic complexities.

REFERENCES

Aligon, J., Golfarelli, M., Marcel, P., Rizzi, S., & Turricchia, E. (2014). Similarity measures for OLAP sessions. Knowledge and Information Systems, 39(2), 463–489. https://doi.org/10.1007/S10115-013-0614-1/METRICS

Berger, B., Waterman, M. S., & Yu, Y. W. (2021). Levenshtein Distance, Sequence Comparison and Biological Database Search. IEEE Transactions on Information Theory, 67(6), 3287–3294. https://doi.org/10.1109/TIT.2020.2996543

Berthelé, R., Lenz, P., & Peyer, E. (2022). Predicting foreign language skills based on first languages: The role of lexical distance and relative morphological complexity. Poznan Studies in Contemporary Linguistics, 58(3), 419–448. https://doi.org/10.1515/PSICL-2022-0020

Bozdag, E. (2013). Bias in algorithmic filtering and personalization. Ethics and Information Technology, 15(3), 209–227. https://doi.org/10.1007/S10676-013-9321-6/METRICS

Bryant, C., Yuan, Z., Qorib, M. R., Cao, H., Ng, H. T., & Briscoe, T. (2023). Grammatical Error Correction: A Survey of the State of the Art. Computational Linguistics, 49(3), 643–701. https://doi.org/10.1162/COLI_A_00478

Chaabi, Y., & Ataa Allah, F. (2022). Amazigh spell checker using Damerau-Levenshtein algorithm and N-gram. Journal of King Saud University - Computer and Information Sciences, 34(8), 6116–6124. https://doi.org/10.1016/J.JKSUCI.2021.07.015

Chen, Y. F., Chocholatý, D., Havlena, V., Holík, L., Lengál, O., & Síč, J. (2023). Solving String Constraints with Lengths by Stabilization. Proceedings of the ACM on Programming Languages, 7(OOPSLA2), 30. https://doi.org/10.1145/3622872

Dashti, S. M. S., Bardsiri, A. K., & Shahbazzadeh, M. J. (2024). Automatic real-word error correction in persian text. Neural Computing and Applications, 1–25. https://doi.org/10.1007/S00521-024-10045-0/METRICS

Espindola, V., Zago, L., Yviquel, H., & Araujo, G. (2023). Source Matching and Rewriting for MLIR Using String-Based Automata. ACM Transactions on Architecture and Code Optimization, 20(2). https://doi.org/10.1145/3571283

Fiscus, J. G., Ajot, J., Radde, N., & Laprun, C. (2006). Multiple Dimension Levenshtein Edit Distance Calculations for Evaluating Automatic Speech Recognition Systems During Simultaneous Speech. LREC. http://www.nist.gov/speech/tools/index.htm

Gou, W., & Chen, Z. (2021). Think Twice: A Post-Processing Approach for the Chinese Spelling Error Correction. Applied Sciences 2021, Vol. 11, Page 5832, 11(13), 5832. https://doi.org/10.3390/APP11135832

Hall, P. A. V., & Dowling, G. R. (1980). Approximate String Matching. ACM Computing Surveys (CSUR), 12(4), 381–402. https://doi.org/10.1145/356827.356830

Janardhana Rao, P., Nageswara Rao, K., Gokuruboyina, S., & Neeraja, K. N. (2024). An Efficient Methodology for Identifying the Similarity Between Languages with Levenshtein Distance. Lecture Notes in Electrical Engineering, 1096, 161–174. https://doi.org/10.1007/978-981-99-7137-4_15

Jongmans, E., Jeannot, F., Liang, L., & Dampérat, M. (2022). Impact of website visual design on user experience and website evaluation: the sequential mediating roles of usability and pleasure. Journal of Marketing Management, 38(17–18), 2078–2113. https://doi.org/10.1080/0267257X.2022.2085315

Khaw, Y. M. J., Tan, T. P., & Bali, R. M. (2024). Hybrid Distance-Statistical-Based Phrase Alignment For Analyzing Parallel Texts In Standard Malay And Malay Dialects. Malaysian Journal of Computer Science, 37(1), 1–25. https://doi.org/10.22452/MJCS.VOL37NO1.5

Khin, D., & Lecturer, P. (2020). International Journal of Advances in Scientific Research and Engineering (ijasre) Similarity Based Information Retrieval Using Levenshtein Distance Algorithm. https://doi.org/10.31695/IJASRE.2020.33780

Kremer, K., & van Manen, S. M. (2023). Design guidelines to improve user experience (UX) in an emergency: On the importance of affordances, signifiers and feedback. Design for Emergency Management, 49–68. https://doi.org/10.4324/9781003306771-4

Kukich, K. (1992). Techniques for automatically correcting words in text. ACM Computing Surveys (CSUR), 24(4), 377–439. https://doi.org/10.1145/146370.146380

Laouafi, A., Laouafi, F., & Boukelia, T. E. (2022). An adaptive hybrid ensemble with pattern similarity analysis and error correction for short-term load forecasting. Applied Energy, 322, 119525. https://doi.org/10.1016

Lau, R. Y. K., Liao, S. Y., Chi-Wai Kwok, R., Xu, K., Xia, Y., & Li, Y. (2012). Text mining and probabilistic language modeling for online review spam detection. ACM Transactions on Management Information Systems (TMIS), 2(4). https://doi.org/10.1145/2070710.2070716

Liu, Y. (2023). Grammatical Error Correction Incorporating First Language Information. https://doi.org/10.25949/23897178.V1

Mashtalir, S. V., Stolbovoi, M. I., & Yakovlev, S. V. (2019). Hybrid Approach to Clustering Various Lengths Video. Journal of Automation and Information Sciences, 51(3), 26–35. https://doi.org/10.1615/JAUTOMATINFSCIEN.V51.I3.30

Maurer, M. E., & Höfer, L. (2012). Sophisticated Phishers Make More Spelling Mistakes: Using URL Similarity against Phishing. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 7672 LNCS, 414–426. https://doi.org/10.1007/978-3-642-35362-8_31

Mehta, A., Salgond, V., Satra, D., & Sharma, N. (2021). Spell Correction And Suggestion Using Levenshtein Distance. *International Research Journal of Engineering and Technology*. www.irjet.net

Ortikov, U. (2023). Practical Uses Of Corpus Analysis In Designing Language teaching materials. Oriental Renaissance: Innovative, Educational, Natural and Social Sciences, 3(7).

Pham, N. L., Vinh Nguyen, V., & Pham, T. V. (2023). A Data Augmentation Method for English-Vietnamese Neural Machine Translation. IEEE Access, 11, 28034–28044. https://doi.org/10.1109/ACCESS.2023.3252898

Ribeiro, L. C., Bernardes, A. T., & Mello, H. (2023). On the fractal patterns of language structures. PLOS ONE, 18(5), e0285630. https://doi.org/10.1371/JOURNAL.PONE.0285630

Schede, E., Brandt, J., Tornede, A., Wever, M., Bengs, V., Hüllermeier, E., & Tierney, K. (2022). A Survey of Methods for Automated Algorithm Configuration. Journal of Artificial Intelligence Research, 75, 425–487. https://doi.org/10.1613/JAIR.1.13676

Skopal, T., & Bustos, B. (2011). On nonmetric similarity search problems in complex domains. ACM Computing Surveys (CSUR), 43(4). https://doi.org/10.1145/1978802.1978813

Suwarningsih, W., & Nuryani. (2024a). Generate fuzzy string-matching to build self attention on Indonesian medical-chatbot. International Journal of Electrical and Computer Engineering, 14(1), 819. https://doi.org/10.11591/IJECE.V14I1.PP819-829

Suwarningsih, W., & Nuryani. (2024b). Generate fuzzy string-matching to build self attention on Indonesian medical-chatbot. International Journal of Electrical and Computer Engineering, 14(1), 819. https://doi.org/10.11591

Szudarski, P. (2023). Collocations, Corpora and Language Learning. Elements in Corpus Linguistics. https://doi.org/10.1017/9781108992602

Walker, S. (2014). Typography & language in everyday life: Prescriptions and practices. Routledge.

Wang, L. L., Cachola, I., Bragg, J., Cheng, E. Y.-Y., Haupt, C., Latzke, M., Kuehl, B., van Zuylen, M., Wagner, L., & Weld, D. S. (2021). Improving the Accessibility of Scientific Documents: Current State, User Needs, and a System Solution to Enhance Scientific PDF Accessibility for Blind and Low Vision Users. https://arxiv.org/abs/2105.00076v1

Wang, Y., Wang, Y., & Liu, Y. (2024). Chinese Spelling Correction Method Based on Multi-feature Fusion and Attention Mechanism. Proceedings of the 3rd International Conference on Computer, Artificial Intelligence and Control Engineering, 481–487. https://doi.org/10.1145/3672758.3672837

Yang, P., Wang, H., Yang, J., Qian, Z., Zhang, Y., & Lin, X. (2024). Deep Learning Approaches for Similarity Computation: A Survey. IEEE Transactions on Knowledge and Data Engineering. https://doi.org/10.1109/TKDE.2024.3422484

Ye, D., Tian, B., Fan, J., Liu, J., Zhou, T., Chen, X., Li, M., & Ma, J. (2023). Improving Query Correction Using Pre-train Language Model In Search Engines. International Conference on Information and Knowledge Management, Proceedings, 2999–3008. https://doi.org/10.1145/3583780.3614930

Young, S. (1997). Corpus-Based Methods in Language and Speech Processing (B. Gerrit, Ed.; Vol. 2). Springer Science & Business Media.

Zhang, S., Hu, Y., & Bian, G. (2017). Research on string similarity algorithm based on Levenshtein Distance. Proceedings of 2017 IEEE 2nd Advanced Information Technology, Electronic and Automation Control Conference, IAEAC 2017, 2247–2251. https://doi.org/10.1109/IAEAC.2017.8054419

Zoya, Latif, S., Latif, R., Majeed, H., & Jamail, N. S. M. (2023). Assessing Urdu Language Processing Tools via Statistical and Outlier Detection Methods on Urdu Tweets. ACM

Transactions on Asian and Low-Resource Language Information Processing, 22(10). https://doi.org/10.1145/3622939