p-ISSN: 2252-3006
e-ISSN: 2685-2411

Implementation of Dictionary Lookup and Damerau-Levenshtein Distance for Correcting Non-Standard Indonesian and English Terms

Dyah Putri Maheswari[a1], Desy Purnami Singgih Putri[a2], Kadek Ayu Wirdiani[a3]
[a]Information Technology Study Program, Faculty of Engineering, Udayana University, Bukit Jimbaran, Bali, Indonesia, Phone. (0361) 701806.
e-mail: [1]pdyah19@gmail.com, [2]desysinggihputri@unud.ac.id, [3]ayuwirdiani@unud.ac.id

Abstrak

Bahasa Indonesia sebagai bahasa resmi mengacu pada Kamus Besar Bahasa Indonesia. Meskipun demikian, dalam penulisan karya ilmiah masih sering ditemukan kesalahan seperti penggunaan kata tidak baku, salah ketik, atau istilah asing yang tidak sesuai konteks. Penelitian ini mengembangkan aplikasi web untuk mendeteksi dan mengoreksi kata tidak baku dalam dokumen ilmiah dengan metode Dictionary Lookup (mengacu KBBI dan istilah TI berbahasa Inggris) dan algoritma Damerau-Levenshtein Distance. Hasil penelitian menunjukkan bahwa sistem yang dikembangkan mampu mendeteksi kata tidak baku dengan tingkat akurasi 100% dan melakukan koreksi dengan precision sebesar 98% serta recall 100% terhadap 100 data uji. Waktu pemrosesan untuk 50 kata baku hanya 0,03 detik, sedangkan untuk 50 kata tidak baku mencapai 57,13 detik. Pemeringkatan berdasarkan frekuensi meningkatkan relevansi saran, dengan koreksi benar sering muncul di peringkat teratas. Jarak edit maksimal yang paling ideal adalah dua, karena masih memberikan hasil yang relevan dan efisien berdasarkan uji penerimaan pengguna (User Acceptance Test).

Kata kunci: *Dictionary Lookup, Damerau-Levenshtein Distance, Deteksi Kata Tidak Baku, Koreksi Kata*

Abstract

Indonesian, as the official language, follows the standards set by the Kamus Besar Bahasa Indonesia. Despite this, academic writing often contains errors such as non-standard word usage, typographical mistakes, and contextually inappropriate foreign terms. This study presents a web-based application for detecting and correcting non-standard words in scholarly documents, employing a Dictionary Lookup approach (referencing both the KBBI and English IT terminology) alongside the Damerau-Levenshtein Distance algorithm. Evaluation on 100 test entries yielded 100% detection accuracy, 98% precision, and 100% recall. Processing times averaged 0.03 seconds for 50 standard words and 57.13 seconds for 50 non-standard words. Incorporating frequency-based ranking improved the relevance of suggestions, with correct corrections most often appearing first. A maximum edit distance of two was identified as optimal, balancing accuracy and efficiency in User Acceptance Testing.

Keywords: *Dictionary Lookup, Damerau-Levenshtein Distance, Non-Standard Word Detection, Word Correction*

1. Introduction

Indonesian language is the official language used throughout Indonesia and forms a vital component of the national identity that must be preserved and developed. In academic writing—such as theses, dissertations, journals, and other scholarly documents—the use of standardized Indonesian is crucial. Adherence to proper linguistic norms not only enhances clarity in communication but also reflects the professionalism of the writer. Employing standard language in scholarly works reduces the potential for misunderstanding and aids readers in better comprehending the author's intent [1].

p-ISSN: 2252-3006
e-ISSN: 2685-2411

Challenges arise for writers as they must ensure that every word conforms to these established standards. One primary reference for verifying standard language is the Kamus Besar Bahasa Indonesia (KBBI). Words that are absent from the KBBI may be deemed non-standard, whether they are foreign terms, misspellings, or typographical errors. This challenge becomes even more significant in fields such as computer science or information technology, where foreign terminology is frequently used [1]

Language errors can be categorized into lapses, errors, and mistakes. Lapses are unintentional omissions or deviations due to shifting one's train of thought before completion. Errors arise from violations of grammatical rules, and mistakes occur when an incorrect word or expression is chosen despite the writer's knowledge of the correct form [2]

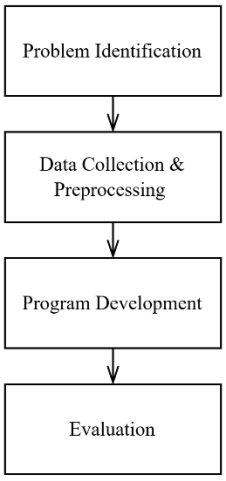

Figure 1. System Development Workflow

Numerous studies have been conducted to identify and analyze errors in the use of Indonesian. For instance, various spelling mistakes have been observed in public signage—such as billboards and institutional nameplates in Palopo—which include issues related to capitalization, abbreviations, punctuation, and non-compliant spellings according to the "Ejaan yang Disempurnakan" [3]. Other research analyzing student academic texts has found that the most common errors occur in letter usage, affixation, punctuation, and the incorporation of foreign words [4]

Two key studies have informed the choice of title and research methods for this study. The first study utilized the Dictionary Lookup method to detect non-standard words based on the KBBI and Kateglo [1]. However, that research only identified non-standard words without offering corrective suggestions. The second study employed Dictionary Lookup to detect spelling errors in documents alongside the Damerau-Levenshtein Distance algorithm to provide correction recommendations [5]. Nevertheless, this study did not consider the usage of terminology specific to particular fields, such as IT-related English terms, and relied solely on edit-distance ranking without further optimization.

Based on these findings, the author has taken the initiative to develop a web-based platform. This platform not only detects non-standard words but also offers corrective recommendations. Moreover, the system is designed to accept inputs containing foreign (English) terminology related to information technology. The system further handles special cases—such as detecting standard words with repeated hyphenation patterns like "anak-anak" or non-repeated patterns like "zig-zag." Fully capitalized abbreviations (e.g., AFNI) and academic titles (e.g., "S.H") are also recognized, with special corrections applied for titles, though errors in titles separated by spaces (e.g., "S H") are not corrected due to tokenization issues. Additionally, a list of frequently misused non-standard words alongside their standard counterparts (e.g., "detil" and "detail") is implemented. Numerical figures and Roman numerals are classified separately so that they are not misidentified as non-standard tokens. The system addresses the use of clitics, prefixes, and suffixes with a predefined subset of particles (namely, "-ku", "ku-", "-mu", "-nya", "ke-", "di-", "-kan", "-an") since the KBBI data already accommodates most other particles such as "ber-" and "peng-

p-ISSN: 2252-3006
e-ISSN: 2685-2411

“ and avoids particles that may undergo fusion or other variations (e.g., “peng-“ transforming into “pe-“, “pem-“, “pen-“, “penge-“, and “peny-“ based on the root word).

The Damerau-Levenshtein algorithm used in this study corrects spelling errors by handling omissions (e.g., “tidakk” for “tidak”), insertions (e.g., “tidk” for “tidak”), substitution errors (e.g., “tidsk” for “tidak”), and transposition errors (e.g., “tidka” for “tidak”) through insertion, deletion, substitution, and transposition operations. Correction candidate ranking is optimized based on the frequency of occurrence in the reference data, ensuring that the most commonly used corrections appear first in the recommendation. Users can directly select and apply the suggested corrections to replace non-standard words.

The Dictionary Lookup method has proven effective in determining whether a word’s spelling is correct or incorrect based on a lexical resource—a database consisting of linguistic data such as corpora, lexicons, word lists, or other forms of language data used for analysis and processing. In addition, the Damerau-Levenshtein Distance algorithm offers superior accuracy in correcting spelling errors compared to the Levenshtein Distance method [5]. The Damerau-Levenshtein algorithm is an enhancement over the Levenshtein Distance algorithm by incorporating an additional transposition operation between two characters, alongside the already existing insertion, deletion, and substitution operations [6]. Studies have shown that this algorithm can address up to 80% of all human spelling errors—including errors such as missing characters, extra characters, or incorrect letter sequences (for instance, “ka” becoming “ak”)—where such errors are counted as two mistakes in the Levenshtein method but only one in the Damerau-Levenshtein approach [7]

The detection process begins with the Dictionary Lookup method, which matches words against all the reference data and specialized lists that have been prepared (including both the KBBI and English IT terminologies). If a word is not found, the system identifies it as an error and subsequently applies the Damerau-Levenshtein Distance algorithm to compute the edit distance and suggest the correction with the smallest distance. The candidate words are then ranked according to their frequency of appearance in the reference data. Through this approach, the application is expected to assist writers in producing scholarly works that adhere to proper Indonesian language standards, while also streamlining the process of correcting non-standard words in academic documents.

2. Research Method / Proposed Method

The research workflow was designed to develop a system for detecting and correcting non-standard words using the Dictionary Lookup method and the Damerau-Levenshtein Distance algorithm. The stages involved include problem identification, data collection and preprocessing, program development, and evaluation.

A. Data Collection & Preprocessing

The data used in the development of this system consists of both primary and secondary data sources. Four main data sources were utilized: the Kamus Besar Bahasa Indonesia (KBBI) Fifth Edition (online version), a list of standard and non-standard word pairs, an English-language glossary of information technology terms, and a word frequency list.

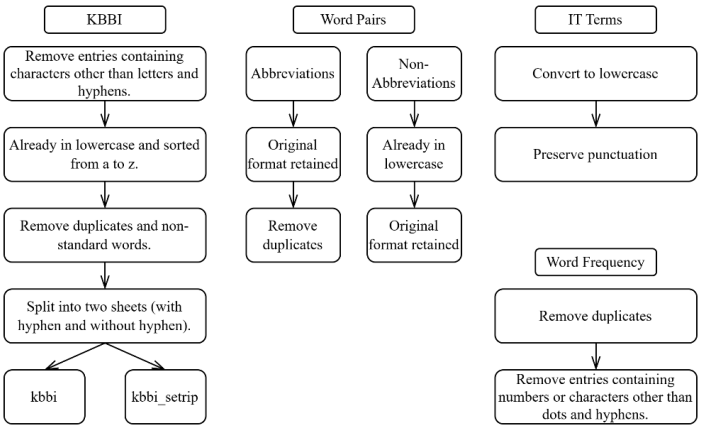

Figure 2. Reference Data Preprocessing

p-ISSN: 2252-3006
e-ISSN: 2685-2411

Primary data refers to data obtained directly from the research subject [8]. The primary data used in this study includes the KBBI, which was retrieved via web scraping from the official online KBBI platform, as well as an English-language information technology glossary compiled from various sources such as Wikipedia and other relevant articles.

Secondary data refers to research data collected indirectly through intermediaries [9]. The secondary data employed in this study includes the list of standard and non-standard word pairs obtained from the Sipebi Open Data repository, and the word frequency list extracted from the Indonesian Corpus. In addition, relevant literature and previous research were also utilized to support and enrich the primary data used in this study.

1) First, the KBBI data underwent a cleaning process to remove entries containing characters other than letters and hyphens. This step ensured that only single-word entries remained, including hyphenated words (e.g., “zig-zag”), as the Dictionary Lookup method operates by matching individual tokens. Entries consisting of only two letters were excluded, as they are deemed irrelevant in the context of KBBI. The KBBI dataset was already in lowercase format and alphabetically ordered from A to Z. Duplicate entries were removed, as well as non-standard words marked with arrow symbols (a default marker from the KBBI source). Finally, the dataset was divided into two separate sheets: one for hyphenated words and another for non-hyphenated words, due to the need for different handling procedures.
2) Second, for the word pair dataset containing abbreviations and academic titles, the original format was preserved—including spacing, capitalization, and punctuation—and duplicate entries were removed.
3) Third, for the dataset containing non-abbreviation word pairs, the data was already in lowercase format, and the original format containing spaces was retained.
4) Fourth, in the IT terminology dataset, all entries were converted to lowercase to normalize capitalization, while punctuation marks such as periods and slashes were preserved.
5) Fifth, in the word frequency dataset, duplicate entries—resulting from repeated downloads—were removed. In addition, entries containing numbers or characters other than periods and hyphens were also excluded.

On the user input side—comprising either text or documents—the preprocessing is not conducted entirely at the outset. Instead, it is carried out in sequential stages tailored to the specific needs of the program.

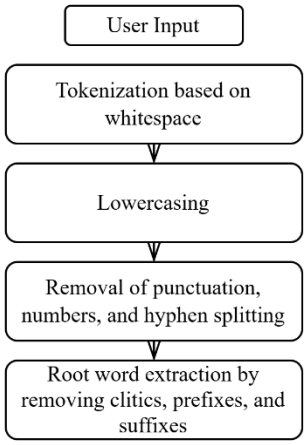

Figure 3. User Input Preprocessing

In the preprocessing of user input, the first step involves tokenization based on whitespace, which splits the input into multiple tokens. However, words containing hyphens (e.g., “anak-anak”) and abbreviated academic titles enclosed with periods (e.g., “S.H.”) are preserved as single tokens. Each token is then analyzed and enters the detection process, with academic titles undergoing a specific format validation. This validation step filters out tokens that do not yet belong to any defined category. These remaining tokens proceed to the next preprocessing stage,

which involves converting all characters to lowercase. A targeted check is then performed to match hyphenated tokens with entries in the reference dataset labeled as "hyphenated data."

Tokens that still do not match any category are passed to the next preprocessing phase, which includes the removal of punctuation and numeric characters, as well as the splitting of hyphenated words. For example, the initial token “anak-anak” is divided into two new tokens: “anak” and “anak”. This step aims to identify whether the split tokens exist individually within the reference data.

The final stage of preprocessing is a limited morphological analysis (affixation), which involves extracting root words by removing specific clitics, prefixes, and suffixes. This process is applied to any remaining unmatched tokens in order to identify possible root forms that may exist in the reference dataset once affixes are removed.

Table 1. User Input Preprocessing Example

Processing	Example
Original Word	Python adalah bahasa pemrograman.
Remove Punctuation	Python adalah bahasa pemrograman
Convert to Lowercase	python adalah bahasa pemrograman
Tokenization	[‘python’, ‘adalah’, ‘bahasa’, ‘pemrograman’]

B. Program Development

The developed system encompasses several functional requirements that must be fulfilled to ensure the effective execution of non-standard word detection and correction processes. The key functional requirements of the system are as follows.

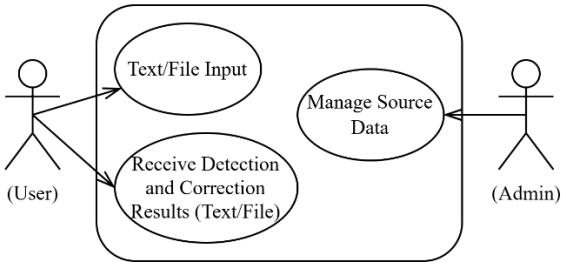

Figure 4. Use Case Diagram

1) The system must be capable of receiving input in the form of either text or files from users. Users can upload document files or directly enter text through the user interface. This input will be processed by the system to detect any non-standard words contained within.
2) The detection process is carried out using the dictionary lookup method, which leverages data from the *Kamus Besar Bahasa Indonesia* (KBBI), a list of standard and non-standard word pairs, and an English dictionary containing information technology terms. The system matches the words in the input text against the entries recognized as standard words within these references.
3) Once non-standard words are identified, the system provides suggestions for their correct standard forms. To determine the most appropriate replacement, the Damerau-Levenshtein Distance algorithm is used to calculate the character-level difference between the non-standard word and available standard words. The system then recommends the word most similar to the detected non-standard word. These candidate words are ranked based on their frequency of occurrence to ensure that the most relevant correction is presented.
4) Administrators play a critical role in maintaining and managing the data utilized by the system. They are responsible for managing various data sources, including updating,

p-ISSN: 2252-3006
e-ISSN: 2685-2411

adding, or removing entries in the KBBI dataset, the list of standard and non-standard word pairs, and the English IT terminology dictionary.

The detection and correction process involves verifying words through a dictionary lookup and, when necessary, applying the Damerau-Levenshtein Distance algorithm to provide correction suggestions. After initial preprocessing, the following sequential checks are performed to classify each word:

1) Abbreviation Pair Matching
   The system checks if a word exists in the abbreviation pair dataset, consisting of non-standard and standard forms. If the word matches the non-standard entry, it is replaced with the corresponding standard form and categorized as “Abbreviation Pair.” If the word matches a standard form, it is considered valid. Additionally, if a word consists entirely of uppercase letters, it is categorized directly as a “Valid Abbreviation.”
2) Hyphenated KBBI Dataset Check
   Words are checked against the KBBI Hyphenated dataset. If found, they are returned with the “KBBI Hyphenated” category.
3) Non-Abbreviation Word Pair Matching
   The system checks for matches in the standard/non-standard word pair dataset (non-abbreviations). If found, the standard word is returned with the category “Word Pair.”
4) Numeric Check
   The system verifies whether the input is a number. If true, it is categorized as “Number.”
5) Roman Numeral Check
   The input is examined using a specific function (is_roman_numeral). If valid, the word is categorized as “Roman Numeral.”
6) KBBI Dataset Check
   The word is validated against the main KBBI dataset. If found, it is returned with the category “KBBI.”
7) IT Terminology Check
   If the word does not match any previous category, the system checks whether it belongs to the IT terminology dataset. If matched, it is returned as “IT Term.”
8) Fallback Affixation Processing
   For words that remain uncategorized, the system performs a final morphological analysis using the periksa_afiksasi function to separate potential clitics, prefixes, or suffixes. If the resulting root word exists in the KBBI dataset, it is classified based on the identified affix type (e.g., clitic, prefix, suffix, or a combination thereof).

Table 2. Dictionary Lookup

Input Word	Category	Result	Explanation
KBBI	Abbreviation	KBBI	Fully capitalized words are considered standard.
IV	Roman Numeral	IV	The word is recognized as a Roman numeral.
Buku	KBBI	Buku	The word is found in the KBBI dataset.
Melanjutkan	KBBI and Clitic	Melanjutkan	The clitic “-kan” is removed; the root “melanjut” is found in the KBBI dataset.
Giji	Word Pair	Gizi	The word is found in the word pair list and corrected to “Gizi.”
Detil	Word Pair	Detail	The word is found in the word pair list and corrected to “Detail.”
Computer	IT Terminology	Computer	The word is found in the information technology terminology dataset.

The Damerau-Levenshtein Distance algorithm is employed to provide correction suggestions for non-standard words by calculating the edit distance between the input word and potential correction candidates. The smaller the edit distance, the higher the candidate word appears in the correction suggestion list. Once the edit distances are computed, the candidates are ranked first by the lowest edit distance, followed by the highest word frequency.

For example, given the user input "Gambarn", the system may generate correction candidates such as "Gambar", "Gambaran", and "Gambir". In cases where multiple candidates

p-ISSN: 2252-3006
e-ISSN: 2685-2411

share the same edit distance, the ranking is determined by the frequency of occurrence in the reference dataset, with more frequently used words—such as "Gambar"—appearing above less common ones like "Gambaran".

Table 3. Damerau-Levenshtein Distance

Input Word	Correction Candidate	Edit Distance	Correction Process
Gambarn	Gambar	1	Deletion of “n”
	Gambaran	1	Insertion of “a”
	Gambir	2	Substitution of “a” with “i”; deletion of “n”

3. Literature Study

A comprehensive literature study is fundamental in establishing a solid theoretical foundation and contextual understanding for the research. This section examines prior studies and methodologies pertinent to the detection and correction of non-standard words, with a particular focus on the Dictionary Lookup method and the Damerau-Levenshtein Distance algorithm.

3.1. Dictionary Lookup

The dictionary lookup method is a straightforward approach designed to identify words within a predefined lexicon. This technique is particularly effective for detecting spelling errors by verifying whether a word exists in a lexical resource [5]. It is commonly employed to identify non-word errors—instances where a misspelled word does not correspond to any entry in the dictionary. The process involves checking if a given word is present in the dictionary; if it is not found, the word is classified as a non-word error [1].

3.2. Damerau-Levenshtein Distance

The Damerau–Levenshtein Distance algorithm extends the Levenshtein Distance by calculating the minimum number of edit operations—insertions, deletions, substitutions, and transpositions—required to transform one string into another [6]. By incorporating the transposition of two adjacent characters as a single operation, this algorithm more accurately models common typographical errors (such as swapped letters) and thereby improves spelling-correction accuracy.

Several studies have demonstrated the advantages of this variant over alternative methods. First, coupling Levenshtein Distance with n-gram ranking based on cosine similarity achieved a precision of 0.97 for insertion errors and a recall of 1.00 for substitution errors [10]. Second, Levenshtein Distance outperformed the Needleman–Wunsch algorithm in both accuracy and execution time, although that comparison did not evaluate the Damerau–Levenshtein variant [11]. Third, a head-to-head comparison of three spelling-correction methods—Peter Norvig’s algorithm (69.09 % accuracy), Levenshtein Distance (73 %), and Damerau–Levenshtein Distance (75 %)—confirmed a two-percentage-point improvement attributable to the character-transposition feature [12]. Fourth, while pure Levenshtein excels at handling simple substitutions and transpositions and n-gram methods are more effective at addressing omissions, the Damerau–Levenshtein variant combines both approaches’ strengths into a single edit-distance framework [13].

In light of these findings, the present study adopts the Damerau–Levenshtein Distance algorithm—augmented by frequency-based candidate ranking—as its primary method for detecting and correcting Indonesian spelling errors, owing to its superior accuracy and efficiency relative to the alternatives.

3.3. Typographical Error

In the process of writing academic papers, typographical errors frequently occur. These errors may result from various factors, including the writer’s limited knowledge of correct spelling according to the Kamus Besar Bahasa Indonesia (KBBI), unintentional carelessness, misconfigured settings in the word-processing application used, or other contributing factors that lead to spelling inaccuracies [14].

Typographical errors can be categorized into two types based on word form: non-word spelling errors and real-word spelling errors. A non-word spelling error refers to a mistake where

p-ISSN: 2252-3006
e-ISSN: 2685-2411

the written word does not exist in the dictionary and has no meaning. In contrast, a real-word spelling error occurs when the word exists in the dictionary and is meaningful, but is not the intended word within the context of the document [5].

3.4. Standard Words in Indonesian

According to the Kamus Besar Bahasa Indonesia (KBBI) Online, Fifth Edition, the term baku is defined as fundamental or primary; a benchmark applied to measure quantity or quality based on consensus; a standard.

A standard word (kata baku) refers to a word that is spoken or written in accordance with established rules or formalized guidelines. These standard rules encompass the Ejaan Yang Disempurnakan (Enhanced Spelling System), standardized grammar, and official dictionaries. Standard words are typically used in formal sentences or in standardized language varieties, whether in spoken or written form [15].

4. Result and Discussion

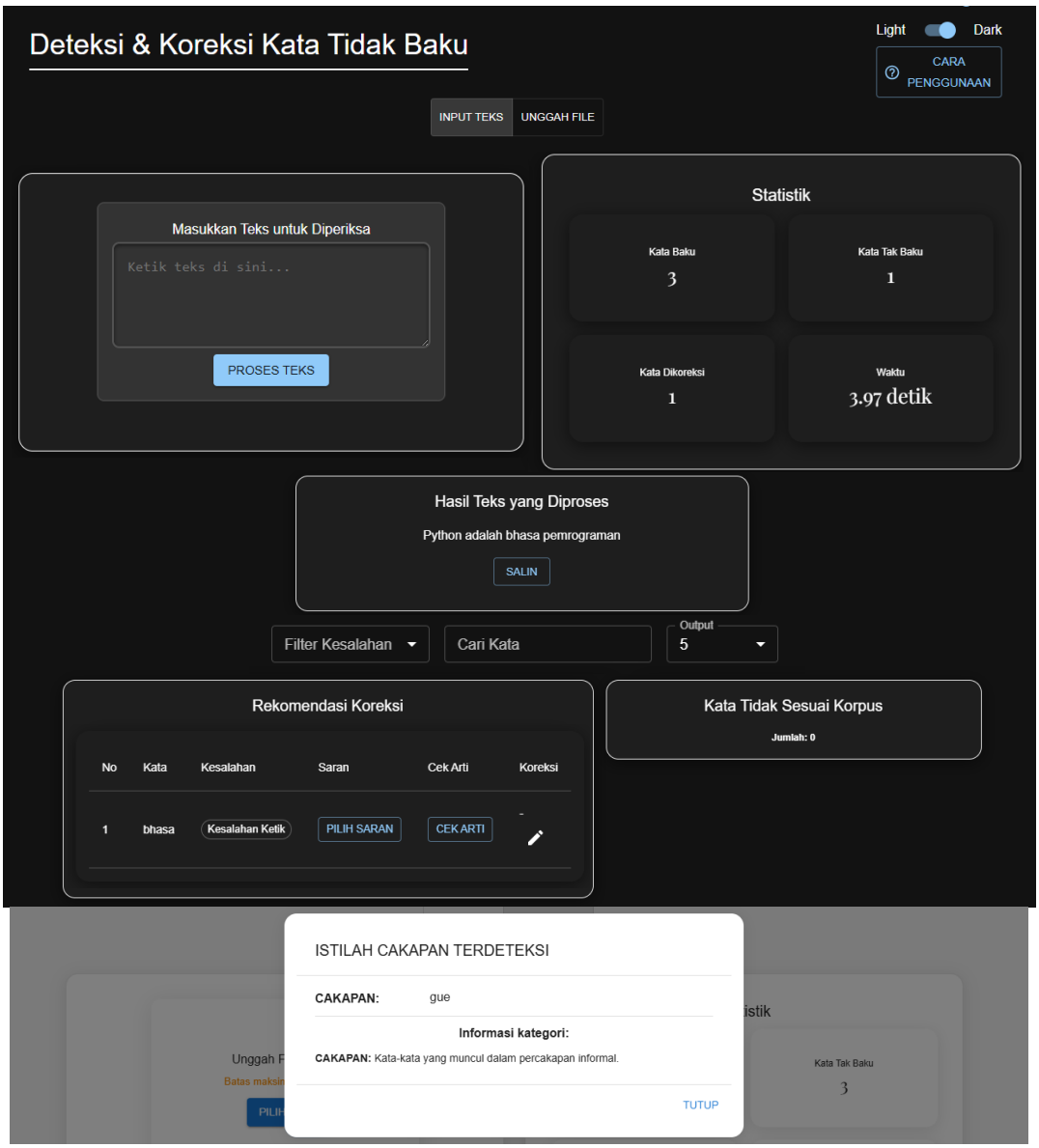

Figure 5. Web Result

This is the web-based interface of the developed detection and correction system. The interface supports both dark and light display modes, as well as two input options: text and file.

p-ISSN: 2252-3006
e-ISSN: 2685-2411

Once the input is processed, the system presents results including: statistics such as the number of standard words, non-standard words, corrected words, and the total processing time.

In addition, the system displays correction recommendations, which include the type of error, selectable correction suggestions, a "check meaning" button, and a manual edit icon. Finally, a list of words not found in the corpus is displayed when the input contains words that exceed the maximum edit distance of two. Moreover, when informal words are detected, a pop-up notification is shown along with a list of the identified informal words.

4.1. Detection Process Testing (Dictionary Lookup)

After testing 50 standard words and 50 non-standard words, the system demonstrated excellent performance in detecting both types. All standard words were correctly identified as standard, and all non-standard words were accurately detected as non-standard. These results indicate that the system made no classification errors, neither false positives nor false negatives.

Thus, the accuracy rate achieved was 100%, reflecting the system's high capability in distinguishing between standard and non-standard words. This success confirms that the Dictionary Lookup approach implemented in the system is highly effective for detecting non-standard words. Given the absence of classification errors, the system can be considered reliable for use in automatic spelling correction applications.

4.2. Correction Process Testing (Damerau-Levenshtein)

The experiment revealed that both frequency-based ranking and non-frequency-based correction yielded similar results: out of 50 test cases, the system provided correct corrections for 49 entries, while 1 entry was incorrectly corrected. However, frequency-based ranking proved more optimal in prioritizing the displayed candidates. It ensured that the correct suggestion appeared as the top candidate, making the correction process more efficient.

The evaluation results yielded a precision of 98% and a recall of 100% for the correction of non-standard words using the Damerau-Levenshtein Distance method, both with and without frequency optimization. These findings indicate a high level of success in generating accurate corrections, despite one instance of incorrect output.

4.3. Processing Time Testing

a. The detection of 50 standard words was completed in a very short time, approximately 0.03 seconds. This result indicates that the standard word classification module performs with high responsiveness and efficiency when processing a dataset of fifty standard words.
b. The detection and correction of 50 non-standard word entries took longer, approximately 57.13 seconds. This increase in processing time was due to the complexity of the process, which involves spelling error detection and the application of the Damerau–Levenshtein algorithm that requires extensive edit distance calculations and candidate ranking.
c. Maximum edit distance beyond 2:
   When the maximum edit distance was increased to three, the processing time significantly rose. The system required 82 seconds to process and correct a single non-standard word. This increase indicates that extending the edit distance threshold directly affects processing duration, as it expands the search space for correction candidates.

4.4. User Acceptance Test (UAT)

User testing was conducted to ensure that the developed application meets users' needs and expectations. A total of 42 students from the Information Technology Study Program at Udayana University tested and evaluated the application. The scenarios and results are as follows all 42 respondents confirmed that the application accurately detects both non-standard words and typographical errors. This aligns with the findings in Section 4.1, where the Dictionary Lookup method achieved 100% detection accuracy. Accordingly, every standard and non-standard word entered into the system was successfully recognized without a single misclassification, underscoring the reliability of the detection module in the context of academic writing.

p-ISSN: 2252-3006
e-ISSN: 2685-2411

Figure 6. Evaluation of Non-Standard Word and Typo Detection Accuracy

Figure 7. Initial Relevance Assessment of Correction Suggestions

A total of 41 out of 42 respondents agreed that the correction suggestions provided by the system were relevant to the intended corrections, while 1 respondent found the suggestion irrelevant. This feedback pertained to the token "ciri has", which was not detected as an error because the Dictionary Lookup module evaluated "ciri" and "has" separately—both of which exist in the KBBI—so the system did not recognize that the phrase as a whole is semantically incorrect. In response, a compound word detection module was added: prior to the lookup process, the system now performs text segmentation to identify token pairs that are semantically invalid as separate units. Based on this enhancement, the input “ciri has” is now recognized as a compound phrase error and is automatically corrected to “ciri khas.”

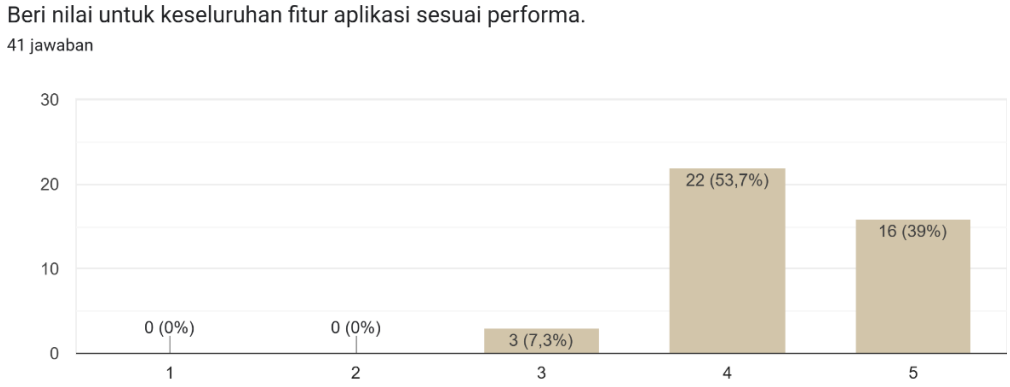

Figure 8. Overall Evaluation of Application Features and Performance

The overall assessment of the application’s features and performance was rated 4.32 out of 5. Sixteen respondents gave a score of 5, twenty-two gave a score of 4, and three respondents gave a score of 3. The predominance of high scores confirms that users are generally satisfied—both functionally and aesthetically—with the application’s features and performance, making it a viable tool for use in academic writing environments.

5. Conclusion

This study successfully developed a web-based application for detecting and correcting non-standard words in scientific documents using the Dictionary Lookup method and the

Damerau-Levenshtein Distance algorithm. The system effectively utilizes the Kamus Besar Bahasa Indonesia (KBBI) and English terminology related to information technology as reference sources to identify non-standard words and provide correction suggestions. Evaluation was conducted on a total of 100 test cases. The detection process achieved an accuracy of 100%, while the correction process yielded a precision of 98% and a recall of 100%.

For 50 standard words tested, the system required only 0.03 seconds of processing time. On the other hand, processing and correcting 50 non-standard words took 57.13 seconds, due to the additional complexity of error detection and correction, which depends on the number of non-standard words identified and the volume of reference data used.

The most ideal maximum edit distance was determined to be 2, as it remained efficient and accurate according to the results of the User Acceptance Test. Furthermore, integrating word frequency into the ranking of correction candidates significantly improved the relevance of suggestions. In testing, the correct word consistently appeared as the top-ranked candidate when frequency-based optimization was applied, compared to when frequency was not used.

The conclusion summarizes the article's main points but does not copy the abstract as a conclusion. A conclusion might emphasize the importance of work results or suggestions for further development.

References

[1] A. Misbullah *et al.*, 'Deteksi Kata Tak Baku dan Kesalahan Penulisan Kata pada Tugas Akhir Mahasiswa Menggunakan Metode Dictionary Lookup', *Jurnal Pendidikan Teknologi informasi*, vol. 6, no. 2, pp. 130–141, 2022.

[2] Jacek Fisiak, *Constrative Linguistics and the Langauge Teacher*. Oxford: Pergamon Press, 1981.

[3] Sukmawaty and Firman, 'Analisis Kesalahan Ejaan Bahasa Indonesia pada Ruang Publik di Kota Palopo'. [Online]. Available: https://sinestesia.pustaka.my.id/journal/article/view/336

[4] B. Dwi Nurwicaksono and D. Amelia, 'Analisis Kesalahan Berbahasa Indonesia pada Teks Ilmiah Mahasiswa', vol. 2, no. 2, 2018, doi: 10.21009/AKSIS.

[5] T. N. Maghfira, I. Cholissodin, and A. W. Widodo, 'Deteksi Kesalahan Ejaan dan Penentuan Rekomendasi Koreksi Kata yang Tepat Pada Dokumen Jurnal JTIIK Menggunakan Dictionary Lookup dan Damerau-Levenshtein Distance', 2017. [Online]. Available: http://j-ptiik.ub.ac.id

[6] J. Jupin, J. Y. Shi, and Z. Obradovic, 'Understanding cloud data using approximate string matching and edit distance', in *Proceedings - 2012 SC Companion: High Performance Computing, Networking Storage and Analysis, SCC 2012*, 2012, pp. 1234–1243. doi: 10.1109/SC.Companion.2012.149.

[7] D. Q. Thang and P. T. Huy, 'Determining Restricted Damerau-Levenshtein Edit-Distance of Two Languages by Extended Automata', *IEEE RIVF International Conference on Computing & Communication Technologies, Research, Innovation, and Vision for the Future (RIVF)*, pp. 1–6, 2010.

[8] J. Rahman, 'Jenis Data Penelitian', 2021.

[9] N. dan B. S. Indriantoro, *Metodologi Penelitian dan Bisnis*. Yogyakarta: BPFE Yogyakarta, 1999.

[10] A. I. Fahma, I. Cholissodin, and R. S. Perdana, 'Identifikasi Kesalahan Penulisan Kata (Typographical Error) pada Dokumen Berbahasa Indonesia Menggunakan Metode N-gram dan Levenshtein Distance', 2018. [Online]. Available: http://j-ptiik.ub.ac.id

[11] Khin Moe Myint Aung, 'Comparison of Levenshtein Distance Algorithm and Needleman–Wunsch Distance Algorithm for String Matching'.

[12] A. Tirta Adi Kusuma and C. Indah Ratnasari, 'Perbandingan Metode Peter Norvig, Levenshtein distance, dan Damerau-Levenshtein distance : Tinjauan Literatur'.

[13] M. Hardiyanti, 'Identifying The Common Type of Spelling Error by Leveraging Levenshtein Distance and N-gram', *Scientific Journal of Informatics*, vol. 8, no. 1, pp. 71–75, May 2021, doi: 10.15294/sji.v8i1.29273.

[14] R. N. Hamzah, 'Aplikasi Perbaikan Ejaan Pada Karya Tulis Ilmiah di Program Studi Teknik Informatika dengan Menerapkan Algoritma Levenshtein Distance', 2016.

[15] E. Kosasih and W. Hermawan, *Bahasa Indonesia Berbasis Kepenulisan Karya Ilmiah dan Jurnal*. Bandung: CV. Thursina, 2012.