ISSN 2723-6471

PRAKE: A Modified RAKE Model for Keyword Extraction in Accreditation Assessment Descriptions

Helena Nurramdhani Irmanda[1], Sri Hartati[2,*], Sri Mulyana[3]

[1]*Doctoral Program in Computer Science, Department of Computer Science and Electronics*

[2,3]*Department of Computer Science and Electronics, Faculty of Mathematics and Natural Sciences, Universitas Gadjah Mada, Yogyakarta 55281, Indonesia*

(Received: November 25, 2025; Revised: February 1, 2026; Accepted: April 12, 2026; Available online: May 3, 2026)

Abstract

Study program accreditation requires aligning assessment criteria with the Self-Evaluation Sheet (LED), which is usually written as a lengthy and complex narrative. Finding relevant information requires a method that can automatically extract keywords from assessment descriptions as representations of the criteria. Keyword extraction can be applied through the Rapid Automatic Keyword Extraction (RAKE) method, a simple technique that works without labeled data. However, standard RAKE uses stopwords as delimiters to segment candidate phrases, making it less effective for complex sentences such as those found in accreditation assessment descriptions. Because a single sentence may contain several ideas, the extraction process should handle phrases carefully through splitting, merging, or extension according to their structure and meaning. To address this limitation, this study introduces PRAKE (Phrase-Refined RAKE), a modified RAKE algorithm that enhances candidate phrase extraction. Modifications are carried out at the Candidate Phrase Extraction stage through three techniques, including Phrase Completion to complete short phrases afterwards with the prefix of the previous phrase, Phrase Restructuring to rearrange phrases through merging or separation based on structure and meaning, and Semantic Phrase Composition to form new phrases from different elements that are semantically interrelated. Additionally, a domain term weighting based on term frequency is integrated into the scoring calculation to strengthen the relevance of terms to the accreditation context. The model achieved a precision of 0.90, recall of 0.83, and F1-score of 0.85, representing the average performance across all 101 assessment descriptions evaluated in this study. The results demonstrate that PRAKE adapts better to accreditation terminology and improves keyword relevance and extraction efficiency. These findings indicate that PRAKE provides a foundation for automated evaluation and can be extended for cross-domain document analysis.

Keywords: Keyword Extraction, Rake, Study Program Accreditation, Natural Language Processing, Text Processing

1. Introduction

The process of accreditation formally evaluates study programs and institutions, using standards defined by the accrediting body [1]. In Indonesia, the Independent Accreditation Institute for Informatics and Computers (LAMINFOKOM) is the authorized body for accrediting study programs in the fields of informatics, computers, and information technology [2]. The accreditation assessment descriptions are typically long, formal narratives explaining the purpose, scope, and indicators of each criterion. Assessors rely on these descriptions to align the information in the Self-Evaluation Sheet (LED) with the relevant criteria. Yet, the narrative nature of these documents often makes the process time-consuming and subjective, particularly because LED texts contain complex sentences that may convey multiple ideas.

Previous studies have explored the use of natural language processing (NLP) in educational and institutional contexts [3], suggesting that automated text analysis could make accreditation review more systematic. NLP, as a branch of artificial intelligence, focuses on how computers can understand and generate human language in textual form [4]. One of its core applications, keyword extraction, aims to identify salient words or phrases that represent the essence of a document's content [5]. Applying this approach to accreditation texts can help highlight essential information more

*Corresponding author: Sri Hartati (shartati@ugm.ac.id)
DOI: https://doi.org/10.47738/jads.v7i2.1057

efficiently and objectively. One widely used method for keyword extraction is Rapid Automatic Keyword Extraction (RAKE) [6]. The algorithm separates text based on stopword boundaries and assigns co-occurrence-based scores [7]. RAKE is lightweight, does not rely on labeled data, and can be applied flexibly across different text types [8]. It has been adopted in many areas, from summarizing software bug reports [9] and extracting keywords from scientific articles [10] to building keyword graphs that visualize large text collections [11]. However, its performance decreases when applied to formal or domain-specific texts. Because RAKE ignores syntactic and semantic structures and depends entirely on the predefined stopword lists, it often generates overly long or fragmented phrases [7] [12].

This limitation makes it less effective in capturing the meaning of complex, narrative documents such as accreditation assessments. Unlike other unsupervised approaches such as YAKE and TextRank, YAKE relies on multiple statistical features and is sensitive to document length and structure [13], while TextRank depends on graph-based co-occurrence that disregards sentence-level structure and deeper semantic relations [14]. In contrast, RAKE focuses on simple phrase-level extraction, making it computationally efficient and easy to refine for domain-specific linguistic adaptation. However, its simplicity also limits the ability to capture deeper contextual and semantic relationships. Therefore, a more adaptive approach that preserves RAKE's efficiency while enhancing its linguistic sensitivity is required to address the contextual nuances of accreditation assessment descriptions.

To address these limitations, this study proposes PRAKE, a tailored modification of the RAKE algorithm for keyword extraction in accreditation assessment descriptions. PRAKE introduces three main enhancements such as refined preprocessing to structure text more effectively, rule-based phrase refinement through completion, restructuring, and semantic composition, and a scoring mechanism that integrates domain-specific term weighting. These components enable PRAKE to better capture the linguistic and contextual nuances of long, formal accreditation narratives. The objective of this study is to demonstrate that PRAKE provides a lightweight yet context-aware alternative for extracting accurate and relevant keywords, thereby supporting assessors in aligning accreditation criteria with the LED.

2. Related Works

Within natural language processing, keyword extraction has been framed in two categories: supervised and unsupervised [15]. The supervised approach in keyword extraction is generally formulated as a binary classification problem, in which phrases in a document are mapped into "keyword" or "non-keyword" classes based on training using labelled data. Machine learning models, from Naïve Bayes, SVM, to deep neural networks such as CNN and RNN have been widely used to perform this classification [16], [17]. As the pre-trained language model evolves, the supervised approach also includes the use of deep learning models such as BERT and RoBERTa, which have shown high performance in extracting key phrases contextually, especially on large-scale data. Nonetheless, this approach has limitations in terms of the need for large amounts of labelled training data as well as high compute resources, which are not always available in every application domain [18].

In contrast, unsupervised approaches do not require labeled data and are often lighter to implement, making them more practical in domains where annotated corpora are unavailable. These unsupervised methods can be grouped into three types: statistical-based, graph-based, and rule-based [19]. Statistical-based methods such as TF-IDF and YAKE use the distribution of words in a document to assess relevance, with YAKE specifically incorporating some local features such as arithmetic position and capitalization frequency to improve extraction results [13]. This approach is relatively simple, but it can result in overly generic keywords if applied to documents with complex structures. Graph-based methods, such as TextRank [20] and PositionRank [21], build a network of relationships between words based on the co-occurrence in a given text window and then calculate the importance score of the word based on the centrality in the graph. Although more context-adaptive than statistical approaches, graph methods generally require more complex processing stages and are not always efficient for long documents or with non-explicit sentence structures [22].

Among the rule-based methods, RAKE is one of the most widely used approaches due to its lightweight nature and non-reliance on training data [6]. RAKE is widely used, but its effectiveness drops in long, domain-specific documents with complex sentence structures [12]. A recent study has explored the use of NLP in accreditation assessments [3]. However, the focus of the research is still limited to content analysis in general and has not explicitly examined keyword extraction as a foundation for automating assessment descriptions. Building on these findings, this study introduces

 ISSN 2723-6471

PRAKE, a modification of RAKE designed to address the challenges of narrative and domain-specific accreditation documents. PRAKE is designed to be more adaptive to the distinctive linguistic structures and terminology of accreditation assessments while preserving the lightweight and unsupervised characteristics of RAKE.

3. Methodology

This study introduces PRAKE, a modified version of the RAKE algorithm aimed at enhancing keyword extraction in formal institutional documents, particularly the adequacy assessment descriptions used in LAMINFOKOM accreditation. These descriptions provide narrative explanations of standards and indicators for evaluating the quality of study programs and are often lengthy and complex. The proposed framework consists of six main components such as data preparation, preprocessing, candidate phrase extraction, score calculation, keyword selection, and testing/evaluation. These components are organized into a sequential yet modular workflow, where each stage produces an output that can serve as the input for the next while remaining independently configurable. As shown in figure 1, the grey blocks represent the standard steps of RAKE, while the yellow blocks highlight the additional steps introduced in PRAKE.

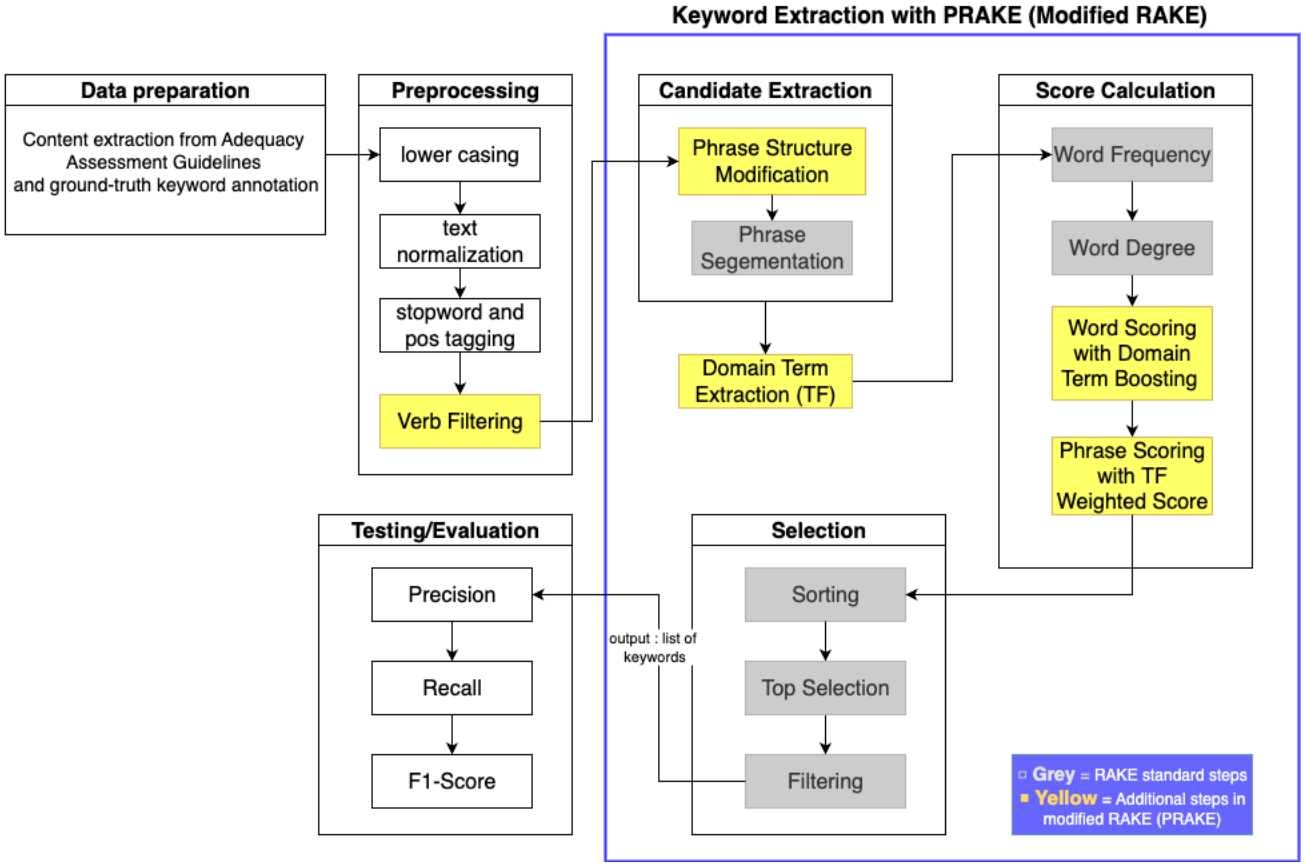

Figure 1. Overall framework of keyword extraction with PRAKE (Modified RAKE)

As illustrated in figure 1, the PRAKE workflow integrates six components that interact programmatically in a structured pipeline. The output of the preprocessing module, which contains clean, tokenized, and POS-tagged text, is used as input for the candidate phrase extraction stage, where linguistic rules are applied to identify and organize candidate phrases. These candidate phrases, together with the domain term list, are then sent to the scoring module that combines RAKE-based and TF-based weights to generate ranked phrase scores. The ranked results are passed to the selection module for filtering and thresholding before being evaluated through precision, recall, and F1-score metrics.

3.1. Data Preparation

The data used in this study were obtained from the Adequacy Assessment Guidelines document published by the Independent Accreditation Institute for Informatics and Computers (LAMINFOKOM). This document contains a

description of standards and indicators that are a reference in assessing the quality of study programs in the field of informatics and computers and are used by assessors in the accreditation process. Initially, the document was compiled in the form of a narrative without an explicit data structure, so a restructuring process was needed so that it could be systematically processed for natural language processing purposes. This process is in line with the data structuring approach commonly performed in domain-specific document processing [23].

The restructuring process is carried out by extracting important content from the document and converting it into a tabular format using an Excel file (.xlsx). The structured dataset consists of 101 rows of data with three main columns: criteria, description, and gt_keyword. The criteria column describes the assessment aspect, the description contains a description of indicators or standards in the form of narrative text, while gt_keyword contains a list of manual keywords as ground truth [24]. The dataset was initially structured and annotated in Excel format to facilitate manual validation by assessors and ease of editing by non-technical contributors. This format was later converted into machine-readable CSV file for integration into the NLP processing pipeline.

The validation of ground truth keywords was conducted by a team of five certified assessors from LAMINFOKOM, each with extensive experience in accreditation evaluation. The assessors collaboratively reviewed and verified the keywords to ensure alignment with the intended assessment indicators and standards. The validation process followed a systematic review procedure, where differences in judgment among assessors were discussed and resolved through joint discussions to reach agreement. This collaborative validation ensured consistent interpretation and reliability of the final ground truth keywords. In addition to verifying keyword alignment, a team of certified LAMINFOKOM assessors also provided qualitative feedback on the readability, completeness, and contextual relevance of the extracted phrases. Their feedback ensured that the resulting keywords and phrases not only aligned with the intended assessment indicators but also maintained clarity and usefulness for practical accreditation evaluation tasks.

In this study, the description column was used as the main input for the keyword extraction process, whereas the gt_keyword column served as the evaluation reference. The keywords in the gt_keyword column were compiled based on the content of the guidelines and had been previously validated to ensure that the ground truth keywords accurately represented the assessment context. Rows with blank values in the description field were removed prior to preprocessing to maintain data integrity. An example of the structured data format is shown in table 1.

Table 1. Example of a structured dataset

Criterion	Description	Gt_keyword
C1.1 PENETAPAN (Establishment): Ketersediaan dokumen kebijakan, standar, IKU, dan IKT yang berkaitan dengan Visi, Misi, Tujuan, Strategi (VMTS) UPPS dan PS (Availability of policy documents, standards, Key Performance Indicators/IKU, and Key Targets/IKT related to the Vision, Mission, Goals, and Strategy/VMTS of the Study Program Unit/UPPS and Study Program/PS)	*Rumusan VMTS UPPS dan PS yang sesuai dengan VMTS PT, memayungi visi keilmuan program studi dan melibatkan pemangku kepentingan internal dan eksternal.* (Formulation of VMTS of UPPS and PS in accordance with the university's VMTS/PT, encompassing the scientific vision of the study program and involving internal and external stakeholders.)	["*rumusan VMTS UPPS*" (VMTS UPPS formulation), "*rumusan VMTS PS*" (VMTS PS formulation), "VMTS PT" (university VMTS), "*visi keilmuan program studi*" (scientific vision of the study program), "*pemangku kepentingan internal*" (internal stakeholders), "*pemangku kepentingan eksternal*" (external stakeholders)]
C1.1 PENETAPAN: Ketersediaan dokumen kebijakan, standar, IKU, dan IKT yang berkaitan dengan Visi, Misi, Tujuan, Strategi (VMTS) UPPS dan PS. (same criterion as above)	*Rumusan strategi pencapaian VMTS UPPS dan PS yang memenuhi tahapan yang jelas, dokumen yang lengkap dan terkait pencapaian visi misi.* (Formulation of VMTS achievement strategies for UPPS and PS that fulfill clear stages, complete documentation, and alignment with the vision and mission.)	["rumusan strategi pencapaian VMTS UPPS" (VMTS UPPS achievement strategy formulation), "strategi pencapaian VMTS PS" (VMTS PS achievement strategy), "tahapan jelas" (clear stages), "dokumen lengkap" (complete document), "pencapaian visi misi" (vision and mission achievement)]
C1.1 PENETAPAN: Ketersediaan dokumen kebijakan, standar, IKU, dan IKT yang berkaitan dengan Visi, Misi, Tujuan, Strategi (VMTS) UPPS dan PS. (same criterion as above)	*Rumusan visi keilmuan PS sesuai KKNI level 6.* (Formulation of the scientific vision of the Study Program in accordance with KKNI/Indonesian National Qualifications Framework level 6.)	["*rumusan visi keilmuan ps*" (scientific vision formulation of the study program), "*KKNI level 6*" (Indonesian National Qualifications Framework level 6)]

3.2. Preprocessing

The preprocessing stage is an important first step in natural language processing to reduce noise and produce a consistent representation of text [25]. In this study, preprocessing was carried out on the description column in the

ISSN 2723-6471

Adequacy Assessment Guidelines document so that it is ready to be processed in the keyword extraction process. There are four main stages of preprocessing carried out for the description column in the Adequacy Assessment Guidelines document, including lowercasing, text normalization, stopword and POS tagging, and verb filtering. The first stage is lowercasing, in which the entire text is converted to lowercase to ensure consistency during the process of word separation and frequency counting [26]. This step also helps avoid word duplication due to capitalization differences. The second stage is text normalization, which includes the removal of non-alphabetic characters, the replacement of symbols (e.g. "&" to "and"), the removal of parentheses and non-essential punctuation, and the equalization of writing formatting to produce clean and uniform text [27].

The third stage is stopword and POS tagging. Stopword tagging is an important step in the preprocessing stage to mark words that are classified as stopwords, i.e., common words with high frequency but low informative value in semantic representations [28]. In this study, the tagging process was carried out using the Sastrawi library, which provides a list of Indonesian stopwords. Words such as "*dari*" (from), "*dengan*" (with), "*yang*" (which/that), and "*untuk*" (for) are marked but not immediately removed, as these markings are necessary to separate the candidate phrases at the keyword extraction stage [29]. In addition, the Part-of-Speech (POS) tagging process was carried out using the Indonesian Stanza model to identify word classes such as nouns, verbs, and adjectives. POS information is used for verb filtering and phrase structure structuring [30].

The fourth stage is verb filtering, in which verbs are excluded from keyword candidates, except for certain verbs that are semantically relevant in the accreditation context, such as "*evaluasi*" (evaluation) and "proses" (process). Verb filtering is carried out based on the results of POS tagging by utilizing the "VERB" label from the Indonesian Stanza model [30]. All verbs identified by the model were excluded from candidate phrases, except for a small subset that frequently appeared in the accreditation corpus and were contextually related to assessment activities. This subset formed the whitelist of semantically relevant verbs, including "*evaluasi*" (evaluation), "proses" (process). By combining POS-tagging with frequency-based selection, the filtering step preserved domain-specific verbs while minimizing the inclusion of generic or context-independent verbs. An example of the text transformation results at each preprocessing stage is shown in table 2.

Table 2. Example of Text Transformation in the Preprocessing Stage

Stages	Output
Original Text	*Rumusan strategi pencapaian VMTS UPPS dan PS yang memenuhi tahapan yang jelas, dokumen yang lengkap dan terkait pencapaian visi misi.* (Formulation of VMTS achievement strategies for UPPS and PS that fulfill clear stages, complete documentation, and alignment with the vision and mission.)
Lowercasing & Normalization	*rumusan strategi pencapaian vmts upps dan ps yang memenuhi tahapan yang jelas dokumen yang lengkap dan terkait pencapaian visi misi*
Stopword Tagging	*rumusan strategi pencapaian vmts upps* [STOP] *ps* [STOP] *memenuhi tahapan* [STOP] *jelas dokumen* [STOP] *lengkap* [STOP] *terkait pencapaian visi misi*
POS Tagging	*rumusan*/NOUN *strategi*/NOUN *pencapaian*/NOUN vmts/PROPN upps/PROPN [STOP] ps/PROPN [STOP] *memenuhi*/VERB *tahapan*/NOUN [STOP] *jelas*/ADJ *dokumen*/NOUN [STOP] *lengkap*/ADJ [STOP] *terkait*/VERB *pencapaian*/NOUN *visi*/NOUN *misi*/NOUN
Verb Filtering	*rumusan strategi pencapaian vmts upps* [STOP] ps [STOP] *tahapan* [STOP] *jelas dokumen* [STOP] *lengkap* [STOP] *pencapaian visi misi*

The result of this preprocessing stage is text that has been segmented, normalized, and cleaned of unimportant words so that it can be used to mark phrase boundaries and define keyword candidates more effectively.

3.3. Candidate Phrase Extraction

After the preprocessing stage is completed, the process continues with candidate phrase extraction. This extraction is carried out through a rule-based approach that consists of two stages, namely modification of phrase structure and phrase segmentation. The first stage, phrase structure modification, aims to refine the phrase structure of the initial segmentation results to make it more representative. There are three main techniques used, including phrase completion, phrase restructuring, and semantic phrase composition. The first technique, phrase completion, adds the

 ISSN 2723-6471

prefix of the previous phrase to the short phrase that appears after it. Without expansion, these short phrases can be taken out of context, reducing the quality of keyword extraction. This technique is applied selectively, especially to phrases included in the whitelist that have been defined based on typical terms in the accreditation domain, such as "PS" (Study Program) or "PT" (Higher Education Institution). For example, if the phrase "ps" appears after "*rumusan VMTS upps*", then the system will form a new phrase "*rumusan VMTS ps*" to maintain the integrity of the meaning. Figure 2 presents the pseudocode of the phrase completion stage, describing the iterative process used to expand short phrases by appending contextual prefixes from preceding phrases.

Phrase Completion

```
Input: Token list with phrase boundaries, whitelist prefix_words
For each phrase after boundary:
    If phrase is short AND in prefix_words:
        Get previous phrase
        Take prefix from previous phrase
        Combine prefix + current phrase
Output: Expanded phrase list
```

Figure 2. Phrase Completion Pseudocode

The second technique, phrase restructuring, restructures phrase structures through merging and splitting. Merges are performed on phrases that semantically form a single entity, such as "*dokumen*" (document) and "*lengkap*" (complete) which are merged into "*dokumen lengkap*" (complete document). Splitting is applied to long phrases containing more than one core concept, for example "*ketersediaan fungsi kelembagaan sistem penjaminan mutu internal*" (availability of institutional function of internal quality assurance system) is separated into "*ketersediaan fungsi kelembagaan*" (availability of institutional function) and "*sistem penjaminan mutu internal*" (internal quality assurance system). To make this process more directed, the system refers to two types of lists, namely a merge whitelist and a split whitelist, each containing pairs of phrases that should be merged or separated. This aims to prevent the processing of semantically irrelevant phrases. Figure 3 presents the pseudocode of the phrase restructuring stage, showing the procedural steps for merging and splitting phrases based on predefined merge and split rules.

Phrase Restructuring

```
Input: Phrase list with <s> separators, merge_rules, split_rules
For each pair of adjacent phrases:
    If pair in merge_rules:
        Merge into one phrase
For each phrase:
    If phrase in split_rules:
        Replace with multiple sub-phrases
Output: Modified phrase list
```

Figure 3. Phrase Restructuring Pseudocode

The third technique, semantic phrase composition, combines scattered elements into new phrases when a semantic relation is detected between a base phrase and a variable attribute or entity. As with other techniques, this expansion refers to a whitelist that specifies the combination of valid root phrases and attributes to combine. For instance, the sentence "*ketersediaan sistem pengelolaan fungsional dan operasional UPPS dan PS*" (availability of functional and operational management systems of UPPS and PS) is expanded into four separate phrases to capture the full semantic meaning hidden in this complex sentence structure. Figure 4 depicts the pseudocode logic of semantic phrase composition, clarifying how base and target phrases are combined into new composite expressions.

Semantic Phrase Composition

```
Input: Phrase list, list of expansion_rules
For each rule in expansion_rules:
    If trigger_phrase found:
        For each base in base_parts:
            For each target in targets:
                Construct: base_prefix + base + target
                Add to phrase list
Output: Combined phrase list
```

Figure 4. Semantic Phrase Composition Pseudocode

The whitelist entries used in phrase completion, restructuring, and semantic composition were derived from frequent co-occurrence patterns and domain terminology observed across the accreditation documents. For example, compound expressions such as "*dokumen lengkap*" (complete document) or "*fungsi kelembagaan*" (institutional function) repeatedly appeared across multiple criteria descriptions and were therefore included as valid merge pairs. Similarly, the split and composition rules were formulated by observing recurring linguistic structures and semantic relationships among phrases in the documents. Although the rule set was handcrafted, its construction was guided by empirical domain patterns and contextual relevance rather than arbitrary selection.

The second stage, phrase segmentation, separates pieces of text into candidate phrases based on the linguistic boundaries that have been marked in the preprocessing [31] stage. In this process, the system uses two main types of markers: (1) the [STOP] tag assigned to the stopword words through the stopword tagging process, and (2) the result of verb filtering, which is a common verb that has been filtered and marked not to be included in the candidate phrase. Both types of markers serve as a phrase cut-off point, and any segment formed between them is considered a candidate phrase. For easy advanced processing, each border between phrases is marked using a special symbol <s>.

Table 3. Examples of before and after phrase segmentation

Before	After Phrase Segmentation
rumusan strategi pencapaian vmts upps <s> rumusan strategi pencapaian vmts ps <s> tahapan jelas, dokumen lengkap <s> pencapaian visi misi <s>	[rumusan strategi pencapaian vmts upps, rumusan strategi pencapaian vmts ps, tahapan jelas, dokumen lengkap, pencapaian visi misi]

3.4. Domain Term Extraction

Domain term extraction is an important stage in the keyword extraction process that aims to identify key terms in the context of program accreditation. The process begins with text preprocessing, including the removal of stopwords using a list from the Sastrawi Library [32] that is manually adjusted to match the characteristics of the accreditation document. After that, term frequency (TF) calculations are carried out to assess how often words appear in the document. To increase relevance, Part-of-Speech (POS) analysis was performed using Stanza to filter word classes [33]. Verbs are generally ignored because they are generic, except for a few words such as "*evaluasi*" (evaluation), "proses" (process), and "*berkala*" (periodic) which are considered important and are still retained. Once the stopwords and common verbs are filtered, the remaining words are sorted by their frequency. Based on an exploratory observation of the word frequency distribution, the top 30% of terms were found to cover most domain-relevant concepts while reducing noise from generic or low-frequency words. Therefore, this threshold was selected as a practical balance between coverage and precision. The resulting list of domain terms is then used in subsequent stages of phrase extraction and scoring to enhance contextual relevance.

3.5. Score Calculation

This score calculation process adapts the basic principles of the RAKE method [6], which calculates word scores based on the ratio between degree and frequency. To increase relevance in the context of the accreditation document, this method is modified with two main additions: (1) the provision of a boosting score on the domain term, and (2) the integration of the RAKE score with the term frequency (TF) weight. This modification aims to make the extraction results more sensitive to important terms in the domain while still maintaining the simplicity of RAKE [34].

Word frequency measures how often a word appears in a candidate's entire set of phrases. The higher the frequency, the more likely it is that the word plays an important role in the topic being discussed [6].

$$\text{Frequency}(w) = \sum_{f \in F_w} \text{the number of occurrences of w in f} \quad (1)$$

w : A word that is evaluated in a phrase. f : The phrase in which the word w appears. F_w : A collection of phrases in which a word w appears. $\text{Frequency}(w)$: The frequency of occurrence of words w in all phrases

Word degree measures the number of other words that appear along with a word in the same phrase. The higher the degree of a word, the more likely it is to be an important link in the phrase structure [6].

 ISSN 2723-6471

$$\text{Degree}(w) = \sum_{f \in F_w} (\text{number of words in f-1}) \quad (2)$$

Degree(w): Degree of a word w, which is the number of other words that appear along with the word w in a phrase

Building on these two measures, word scoring with domain term boosting calculates the initial score of a word based on the ratio between degree and frequency according to the RAKE principle. To increase sensitivity to important terms in the context of accreditation, words included in the list of domain terms are given a boosting multiplier. The boost is worth more than 1 if the word belongs to the term domain and 1 if it does not.

$$\text{Word Score}(w) = \frac{\text{Degree}(w)}{\text{Frequency}(w)} \; x \; boost \quad (3)$$

Finally, the phrase score with TF-weighted scoring. A phrase score is calculated by summing up all the scores of its forming words. To increase contextual relevance, the weight of term frequency (TF) is added as a second factor. The final score is calculated by a linear combination of the two scores.

$$\text{Phrase Score}(P) = \alpha \cdot \sum_{i=1}^{n} \left(\frac{\text{Degree}(w_i)}{\text{Frequency}(w_i)} \right) + \beta \cdot \sum_{i=1}^{n} \text{TF}(w_i) \quad (4)$$

P: Candidate Phrase. w_i : the i-word in the phrase. α, β : weights to set the contribution of each score (α+β=1). In this study, a α of 0.7 and β of 0.3 were used to combine the RAKE score and the TF domain weight linearly together.

The weighting ratio of α = 0.7 and β = 0.3 was determined through exploratory tuning to balance the contribution of RAKE's structural co-occurrence score and term frequency weighting. Several proportional combinations (e.g., 0.5–0.5, 0.6–0.4, and 0.7–0.3) were observed, and the chosen ratio produced the most consistent and contextually coherent extraction results. This configuration emphasizes structural relationships between words while maintaining contextual relevance through frequency weighting, making it well-suited for narrative accreditation texts.

3.6. Selection and Filtering

Selection and filtering aim to filter and select the best phrases from a group of candidates who have gone through the score calculation process. First, in the sorting stage, after each candidate's phrase obtains a combined score, all phrases are sorted from highest to lowest score. This step is done so that the phrase with the greatest contribution of meaning to the content of the document is at the very top, facilitating the next selection process. Next, in the top selection stage, phrases whose values are above the median are selected to ensure that only phrases with the most significant semantic contribution are taken as prime keyword candidates. To avoid too few results, the system also sets a minimum limit of the top five phrases to maintain the usefulness of the extraction results. Finally, in the filtering stage, selected phrases are filtered to remove duplicates or meaningless phrases, such as very short phrases or those consisting of only a single common word. Only phrases that pass this stage are included in the final list of extracted keywords.

This selection and screening step is the final determinant of the quality of the extraction results, because it ensures that only phrases that are informative, relevant, and representative of the content of the document are used as keywords. The use of the median as a threshold aims to balance the trade-off between precision and recall by retaining phrases with above-average semantic contribution while filtering out low-relevance candidates. To avoid losing infrequent but contextually significant terms, the system also enforces a minimum selection rule that includes the top five phrases regardless of score distribution. This approach helps maintain coverage of meaningful but less frequent expressions, reducing potential bias toward common phrases.

3.7. Evaluation Metrics

In the keyword extraction task, the precision, recall, and F1-score metrics are used to evaluate the model's performance in recognizing relevant keywords. Precision measures the accuracy of the model in recognizing the essential elements of the overall elements detected by the model. The higher the precision value, the fewer incorrect elements are known as essential elements. Recall measures the extent to which the model is able to recognize all the important elements of

 ISSN 2723-6471

the data. A high recall value indicates that the model can capture most of the relevant elements. While the F1-score is a harmonic mean between precision and recall, it gives an overall picture of the model's performance [35].

Keyword extraction testing was conducted to assess the model's ability to recognize important keywords from adequacy assessment guidelines. The model was evaluated using the precision, recall, and F1-score metrics that have been described in Equation 5, Equation 6, and Equation 7 [36]. The ground truth consisted of manually annotated keywords that served as reference labels. In this evaluation, True Positives (TP) represent the number of relevant keywords correctly identified by the model, False Positives (FP) denote irrelevant words incorrectly identified as keywords, and False Negatives (FN) refer to relevant keywords that the model failed to detect [37].

$$Precision = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Positives (FP)}} \tag{5}$$

$$\text{Recall} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}} \tag{6}$$

$$\text{F1-Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{7}$$

True Positives (TP): relevant keywords correctly identified by the model. False Positives (FP): non-relevant words incorrectly identified as keywords. False Negatives (FN): relevant keywords that were not detected by the model.

The evaluation was performed by applying the model to the text of the adequacy assessment guidelines to extract keywords relevant to each assessment criterion. Each extracted keyword was then compared with the ground truth to measure how accurately the model identified relevant keywords.

To ensure fair performance comparison across all samples, the evaluation metrics in this study were computed using the macro-averaging approach. Precision, recall, and F1-score were first calculated for each individual description and then averaged across all 101 data samples. This approach was chosen because each description represents an independent accreditation criterion with a different number of reference keywords. Macro-averaging therefore provides a balanced view of the model's overall consistency without being disproportionately influenced by longer descriptions containing more keywords.

In addition to the standard evaluation metrics, a statistical significance analysis was conducted as an additional step in the evaluation phase to verify that the observed improvements were not due to random variation. To confirm the statistical significance of the performance difference between PRAKE and standard RAKE, a Wilcoxon signed-rank test was applied to the paired F1-scores from all 101 samples [38]. The Shapiro–Wilk test, which provides a more reliable normality assessment for small- to medium-sized datasets [39], indicated that the data were not normally distributed ($p < 0.05$), making the non-parametric Wilcoxon approach more suitable for evaluating paired performance results.

4. Results and Discussion

This section presents the evaluation results of PRAKE, the proposed modification of the RAKE model. Evaluation was carried out by comparing the extraction results obtained using this method with the results of the standard RAKE model. The evaluation process was carried out on 101 description data derived from the Adequacy Assessment Guidelines document, using the ground truth keywords as a reference. In addition to metric-based quantitative evaluations (precision, recall, and F1-score), word cloud visualizations are also used to provide an overview of the distribution and dominance of terms in data sets. The word cloud of top domain terms can be seen in figure 5.

ISSN 2723-6471

Figure 5. Overview of wordcloud of top domain terms (30%)

Figure 5 is a word cloud visualization that displays the top 30% of domain terms extracted based on term frequency after going through data cleaning processes such as stopword removal and general verb exclusion. This visualization aims to identify the most prominent and representative terms in the description of the assessment of accreditation assessment items. Words with larger sizes indicate a higher frequency, for example: "*pelaksanaan*" (implementation), "*proses*" (process), "*penelitian*" (research), "*standar*" (standard), "*evaluasi*" (evaluation), and "*ketersediaan*" (availability), which represent key themes in the assessment narrative. This indicates a strong assessment focus on aspects of program implementation, standard fulfillment, and support for the *tridharma* (three pillars of higher education: teaching, research, and community service) of higher education. In addition, some smaller words such as "*penyesuaian*" (adjustment), "*strategi*" (strategy), "*dokumen*" (document), and "*pengelolaan*" (management) still appear as part of the relevant domain terms, albeit with a lower frequency.

The word cloud visualization in figure 5 mainly serves an exploratory and illustrative function rather than an analytical one. Its purpose is to provide a visual impression of term prominence and thematic distribution within the collection of accreditation descriptions, offering preliminary insights before the quantitative evaluation. The subsequent analysis focuses on the quantitative assessment of PRAKE's performance using precision, recall, and F1-score metrics. Each extracted keyword was compared with the manually prepared ground truth to measure accuracy in recognizing relevant keywords. Table 4 presents sample evaluation results showing the correspondence between the extracted and reference keywords along with their respective precision, recall, and F1-score values.

Table 4. Sample Evaluation per Line using the Proposed Method (PRAKE)

Description	Extracted Keyword From PRAKE	Ground Truth Keyword	Precision	Recall	F1-Score
Rumusan VMTS UPPS dan PS yang sesuai dengan VMTS PT, memayungi visi keilmuan program studi dan melibatkan pemangku kepentingan internal dan eksternal. (Formulation of VMTS of UPPS and PS in accordance with the university's VMTS, encompassing the scientific vision of the study program and involving internal and external stakeholders.)	*visi keilmuan program studi, pemangku kepentingan eksternal, pemangku kepentingan internal, rumusan vmts ps, rumusan vmts upps* (Scientific vision of the study program, external stakeholders, internal stakeholders, formulation of VMTS of the study program, formulation of VMTS of UPPS)	*rumusan vmts upps, rumusan vmts ps, vmts pt, visi keilmuan program studi, pemangku kepentingan internal, pemangku kepentingan eksternal* (Formulation of VMTS of UPPS, formulation of VMTS of the study program, VMTS of the university, scientific vision of the study program, internal stakeholders, external stakeholders)	1	0.83	0.91
Rumusan strategi pencapaian VMTS UPPS dan PS yang memenuhi tahapan yang jelas, dokumen yang lengkap dan terkait pencapaian visi misi. (Formulation of VMTS achievement strategies for UPPS and PS that fulfill clear stages, complete documentation, and alignment with the vision and mission.)	*dokumen lengkap, pencapaian visi misi, rumusan strategi pencapaian vmts ps, rumusan strategi pencapaian vmts upps, tahapan jelas* (Complete documents, achievement of vision and mission, formulation of strategies for achieving VMTS of the study program, formulation of strategies for achieving VMTS of UPPS, clear stages)	*rumusan strategi pencapaian vmts upps, strategi pencapaian vmts ps, tahapan jelas, dokumen lengkap, pencapaian visi misi* (Formulation of strategies for achieving VMTS of UPPS, strategies for achieving VMTS of the study program, clear stages, complete documents, achievement of vision and mission)	1	1	1
Rumusan visi keilmuan PS sesuai KKNI level 6. (Formulation of the scientific vision of the Study Program in accordance with KKNI/Indonesian National Qualifications Framework level 6.)	*kkni level 6, rumusan visi keilmuan ps* (KKNI level 6, formulation of the scientific vision of the study program)	*rumusan visi keilmuan ps, kkni level 6* (Formulation of the scientific vision of the study program, KKNI level 6)	1	1	1
Keterlaksanaan VMTS UPPS dan PS yang sesuai dengan VMTS PT, memayungi visi keilmuan Program Studi dan melibatkan pemangku kepentingan internal dan eksternal. (Implementation of VMTS of UPPS and PS in accordance with the university's VMTS, encompassing the scientific vision of the Study Program and involving internal and external stakeholders.)	*visi keilmuan program studi, pemangku kepentingan eksternal, pemangku kepentingan internal, keterlaksanaan vmts ps, keterlaksanaan vmts upps* (Scientific vision of the study program, external stakeholders, internal stakeholders, implementation of VMTS of the study program, implementation of VMTS of UPPS)	*keterlaksanaan vmts upps, keterlaksanaan vmts ps,vmts pt, visi keilmuan program studi, pemangku kepentingan internal, pemangku kepentingan eksternal* (Implementation of VMTS of UPPS, implementation of VMTS of the study program, VMTS of the university, scientific vision of the study program, internal stakeholders, external stakeholders)	1	0.83	0.91

ISSN 2723-6471

After evaluating each record, the precision, recall, and F1-score values were averaged to obtain the overall performance of the proposed method. Table 5 summarizes the comparison between PRAKE and the standard RAKE model.

Table 5. Model Evaluation Average

Model	Precision	Recall	F1-Score
PRAKE (proposed method)	0.90	0.83	0.85
Standard RAKE	0.59	0.67	0.62

From table 5, it can be seen that the PRAKE consistently shows better performance than the standard RAKE, with a precision value of 0.90, a recall of 0.83, and an F1-score of 0.85. In comparison, standard RAKE obtained a precision of 0.59, a recall of 0.67, and an F1-score of 0.62. These results show that the modifications made to the RAKE algorithm are able to significantly improve both precision and recall, which is reflected in an increase in F1-score of almost 10 points compared to the standard version. To clarify the performance comparison between the models, figure 6 shows the precision, recall, and F1-score values of the two approaches in the form of a bar graph.

To ensure that the observed improvement in performance was not due to random variation, a statistical significance analysis was performed using the Wilcoxon signed-rank test on the paired F1-scores from all 101 evaluation samples. The test yielded a test statistic of $W = 3390.0$ with a p-value < 0.01, confirming that the improvement achieved by PRAKE over the standard RAKE is statistically significant at the 99% confidence level. This result demonstrates that the enhancements introduced in PRAKE, particularly the phrase restructuring process and the domain term weighting mechanism, contribute to a consistent and reliable improvement rather than a random effect. As illustrated in figure 6, PRAKE consistently achieves higher precision, recall, and F1-score compared with the standard RAKE.

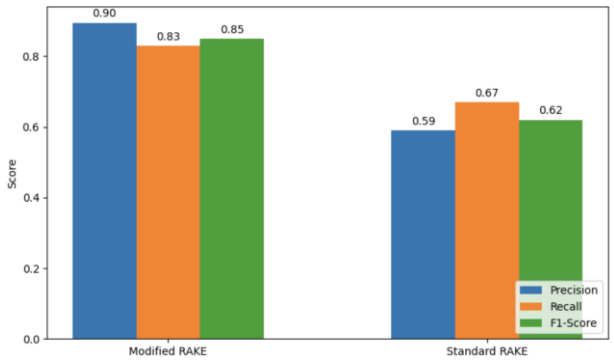

Figure 6. Comparison Chart of Standard RAKE and PRAKE Evaluation Values

As shown in figure 6, PRAKE performs better than the standard RAKE across all evaluation metrics. The most notable improvement is in precision, which means that the phrases generated by PRAKE match the ground truth keywords more accurately. This improvement is mainly the result of PRAKE's ability to adjust phrase boundaries and apply domain-specific term weighting, so irrelevant fragments are reduced and incomplete phrases are avoided. In addition, the balanced rise in recall and F1-score shows that PRAKE can capture a broader range of relevant keywords while still maintaining high precision. Taken together, these results suggest that the modifications in PRAKE, particularly phrase restructuring and semantic composition, make it more effective for handling accreditation documents that are typically long and filled with domain-specific terminology. To further illustrate this improvement, table 6 presents an example of keyword extraction results from a single line of description, comparing PRAKE with the standard RAKE.

Table 6. Example of Extraction Results

Description	*PRAKE*	*Standard RAKE*	*Ground Truth Keyword*
Rumusan VMTS UPPS dan PS yang sesuai dengan VMTS PT, memayungi visi keilmuan program studi dan melibatkan pemangku kepentingan internal dan eksternal.	*visi keilmuan program studi, pemangku kepentingan eksternal, pemangku kepentingan internal, rumusan vmts ps, rumusan vmts upps*	*memayungi visi keilmuan program studi, melibatkan pemangku kepentingan internal, rumusan vmts upps, vmts pt, ps, eksternal*	*rumusan vmts upps, rumusan vmts ps, vmts pt, visi keilmuan program studi, pemangku kepentingan internal, pemangku kepentingan eksternal*

ISSN 2723-6471

(Formulation of VMTS of UPPS and PS in accordance with the university's VMTS, encompassing the scientific vision of the study program and involving internal and external stakeholders.)			
Rumusan strategi pencapaian VMTS UPPS dan PS yang memenuhi tahapan yang jelas, dokumen yang lengkap dan terkait pencapaian visi misi. (Formulation of VMTS achievement strategies for UPPS and PS that fulfill clear stages, complete documentation, and alignment with the vision and mission.)	*dokumen lengkap, pencapaian visi misi, rumusan strategi pencapaian vmts ps, rumusan strategi pencapaian vmts upps, tahapan jelas*	*rumusan strategi pencapaian vmts upps, terkait pencapaian visi misi, memenuhi tahapan, ps, jelas, dokumen, lengkap*	*rumusan strategi pencapaian vmts upps, rumusan strategi pencapaian vmts ps, tahapan jelas, dokumen lengkap, pencapaian visi misi*
Rumusan visi keilmuan PS sesuai KKNI level 6. (Formulation of the scientific vision of the Study Program in accordance with KKNI/Indonesian National Qualifications Framework level 6.)	*kkni level 6, rumusan visi keilmuan ps*	*rumusan visi keilmuan ps, kkni level 6*	*rumusan visi keilmuan ps,kkni level 6*
Keterlaksanaan VMTS UPPS dan PS yang sesuai dengan VMTS PT, memayungi visi keilmuan Program Studi dan melibatkan pemangku kepentingan internal dan eksternal. (Implementation of VMTS of UPPS and PS in accordance with the university's VMTS, encompassing the scientific vision of the Study Program and involving internal and external stakeholders.)	*visi keilmuan program studi, pemangku kepentingan eksternal, pemangku kepentingan internal, keterlaksanaan vmts ps, keterlaksanaan vmts upps*	*memayungi visi keilmuan program studi, melibatkan pemangku kepentingan internal, keterlaksanaan vmts upps, vmts pt, ps, eksternal*	*keterlaksanaan vmts upps, keterlaksanaan vmts ps,vmts pt, visi keilmuan program studi, pemangku kepentingan internal, pemangku kepentingan eksternal*

PRAKE is able to capture a more complete phrase structure, while the standard RAKE model often breaks phrases into separate units of words. This shows the advantages of the context extension and phrase control approach at the preprocessing and extraction candidate stages. The results of the experiment showed a number of important findings that confirmed the effectiveness of the modifications made to the RAKE algorithm. First, the addition of a boosting mechanism to the term domain combined with weighting based on term frequency was observed to improve the relevance of keyword extraction results. Words that often appear in the context of the accreditation domain, such as "*strategi*" (strategy), "*kebijakan*" (policy), "*penelitian*" (research), and "*pengabdian masyarakat*" (community service), earn higher scores because they're given additional weight. This makes the system more sensitive to semantically important terms in the accreditation guideline.

In addition, the use of POS-tagging-based phrase separation as well as special rules based on sentence structure makes a clear contribution to forming more precise phrase candidates. Unlike standard RAKE, which separates phrases based only on stopwords, this approach is able to recognize the natural limits of phrases that are more syntactically and semantically accurate. For example, phrases such as "*dokumen pengelolaan pkm lengkap*" (complete community service management document) or "*monitoring kesesuaian penelitian mahasiswa*" (monitoring of student research conformity) were recognized as meaningful units, rather than split into less informative parts.

Another important finding was the emergence of long phrases (three to five words) among the top-ranked keywords. Under the standard RAKE, such phrases often received low scores or failed to appear at all, since the algorithm tends to prioritize shorter candidates. By modifying the scoring mechanism, PRAKE became more tolerant of longer, information-rich phrases particularly when they contain domain-relevant terms. According to the assessors' qualitative feedback, the phrases generated by PRAKE were generally clearer, more informative, and more contextually relevant for accreditation evaluation than those produced by standard RAKE. While some of these phrases were longer, the assessors noted that the added length contributed to greater completeness and clarity rather than redundancy. Overall, these findings indicate that the proposed approach improves not only quantitative performance (precision, recall, and F1-score), but also the qualitative aspect of the extracted keywords. The extracted keywords are more representative, contextual, and reliable to support assessors in matching information within the LED. Nevertheless, PRAKE remains a rule-based method, which means its effectiveness still depends on handcrafted rules and domain-specific weighting. Future work may focus on refining these rules or combining PRAKE with machine learning approaches to achieve greater adaptability.

5. Conclusion

This study proposes PRAKE, a modified version of the RAKE algorithm, to improve keyword extraction in the LAMINFOKOM Adequacy Assessment Guidelines document. PRAKE incorporates POS tagging-based verb filtering, rule-based phrase segmentation, and domain-specific term weighting using term frequency (TF). This approach addresses the weaknesses of the RAKE standard in extracting long phrases and relevant terms in the context of accreditation. The results of the evaluation showed an increase in performance with a precision value of 0.90, a recall

of 0.83, and an F1-score of 0.85. The resulting phrases are also more representative and in accordance with the accreditation terminology. Overall, PRAKE contributes to the automation of keyword extraction, particularly for matching the scoring criteria and the content of the LED.

Future work may focus on refining these rules and developing a hybrid model that integrates PRAKE's rule-based framework with embedding-based methods such as BERT to enhance semantic similarity measurement and synonym expansion, thereby improving the model's overall performance on keyword extraction tasks. In addition, mitigating PRAKE's reliance on handcrafted rules and fixed thresholds through automatic rule tuning, adaptive threshold optimization, or embedding-based contextual weighting could further improve its robustness. Furthermore, although this study focuses on the accreditation domain, PRAKE's modular design allows adaptation to other text domains by adjusting the domain lexicon and phrase segmentation rules. This adaptability will be explored in future cross-domain validation studies to evaluate the model's generalizability and scalability.

6. Declarations

6.1. Author Contributions

Conceptualization: HNI, SH; Methodology: HNI, SH, SM; Software: HNI; Validation: SH, SM; Formal Analysis: HNI; Investigation: HNI; Resources: SH; Data Curation: HNI; Writing Original Draft Preparation: HNI; Writing Review and Editing: HNI, SH, SM; Visualization: HNI; Supervision: SH, SM. All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

6.3. Funding

The authors received no financial support for the research, authorship, and/or publication of this article

6.4. Institutional Review Board Statement

Not applicable.

6.5. Informed Consent Statement

Not applicable.

6.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] L. Sudianto and P. Simon, "Application of monitoring database for accreditation instrument UKI PAULUS," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 846, no. May, pp. 12027–12034, 2020, doi: 10.1088/1757-899X/846/1/012027.

[2] LAMINFOKOM, "Academic Paper for Study Program Accreditation," 2021. [Online]. Available: https://laminfokom.or.id/official/img/instrumen/instrumen_2372Lampiran%201%20PerBAN-PT%2015%202021%20Instrumen%20APS%20Sarjana%20Infokom.pdf. [Accessed: Feb. 28, 2024].

[3] A. Mulyanto, S. Hartati, and R. Wardoyo, "An integrated model of natural language processing technique and case-based reasoning for supporting study program accreditation," *ICIC Express Lett.*, vol. 18, no. Jul., pp. 749–757, 2024, doi: 10.24507/icicel.18.07.749.

[4] S. Mulyana, S. Hartati, R. Wardoyo, and Subandi, "A processing model using natural language processing (NLP) for narrative text of medical record for producing symptoms of mental disorders," *Int. J. Inform. Comput.*, vol. 2019, no. Sep., pp. 1–6, 2019, doi: 10.1109/ICIC47613.2019.8985862.

[5] Z. H. Amur, Y. K. Hooi, G. M. Soomro, H. Bhanbhro, S. Karyem, and N. Sohu, "Unlocking the potential of keyword extraction: the need for access to high-quality datasets," *Appl. Sci.,* vol. 13, no. Jun., pp. 7228–7235, 2023, doi: 10.3390/app13127228.

[6] S. Rose, D. Engel, N. Cramer, and W. Cowley, "Automatic keyword extraction from individual documents," *Text Min. Appl. Theory*, vol. 2010, no. Jan., pp. 1–20, 2010, doi: 10.1002/9780470689646.ch1.

[7] M. Nadim, D. Akopian, and A. Matamoros, "A comparative assessment of unsupervised keyword extraction tools," *J. Inf. Sci.,* vol. 49, no. Apr., pp. 573–588, 2023, doi: 10.1109/ACCESS.2023.3344032.

[8] V. Singh and B. K. Bolla, "Hybrid approach to unsupervised keyphrase extraction," *Procedia Comput. Sci.*, vol. 235, no. Jan., pp. 1498–1511, 2024, doi: 10.1016/j.procs.2024.04.141.

[9] S. G. Jindal and A. Kaur, "Automatic keyword and sentence-based text summarization for software bug reports," *IEEE Access,* vol. 8, no. Jan., pp. 65352–65370, 2020, doi: 10.1109/ACCESS.2020.2985222.

[10] K. Rinartha and L. G. S. Kartika, "Rapid automatic keyword extraction and word frequency in scientific article keywords extraction," *Int. J. Cybern. Intell. Syst*., vol. 2021, no. Oct., pp. 1–4, 2021, doi: 10.1109/ICORIS52787.2021.9649458.

[11] G. Muppala and T. Devi, "Accurate recasting of giant text into charts using rapid automatic keyword extraction algorithm in comparison with bag of words algorithm," *Int. J. Contemp. Comput. Inform*., vol. 2023, no. Sep., pp. 2548–2552, 2023, doi: 10.1109/IC3I59117.2023.10397804.

[12] E. Papagiannopoulou and G. Tsoumakas, "A review of keyphrase extraction," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.,* vol. 10, no. Mar., pp. e1339–e1345, 2020, doi: 10.1002/widm.1339.

[13] R. Campos, V. Mangaravite, A. Pasquali, A. Jorge, C. Nunes, and A. Jatowt, "YAKE! keyword extraction from single documents using multiple local features," *Inf. Sci.*, vol. 509, no. Jan., pp. 257–289, 2020, doi: 10.1016/j.ins.2019.09.013.

[14] R. Mihalcea and P. Tarau, "Textrank: bringing order into text," *Empir. Methods Nat. Lang. Process.*, vol. 2004, no. Jan., pp. 404–411, 2004.

[15] N. Firoozeh, A. Nazarenko, F. Alizon, and B. Daille, "Keyword extraction: issues and methods," *Nat. Lang. Eng*., vol. 26, no. May, pp. 259–291, 2020, doi: 10.1017/S1351324919000457.

[16] L. Ajallouda, F. Z. Fagroud, A. Zellou, and E. H. Benlahmar, "Automatic keyphrases extraction: an overview of deep learning approaches," *Bull. Electr. Eng. Inform*., vol. 12, no. Jan., pp. 303–313, 2023, doi: 10.11591/eei.v12i1.4130.

[17] Z. Alami Merrouni, B. Frikh, and B. Ouhbi, "Automatic keyphrase extraction: a survey and trends," *J. Intell. Inf. Syst*., vol. 54, no. Feb., pp. 391–424, 2020, doi: 10.1007/s10844-019-00558-9.

[18] B. Xie, "From statistical methods to deep learning, automatic keyphrase prediction: a survey," *Inf. Process. Manage*., vol. 60, no. Apr., pp. 103382–103390, 2023, doi: 10.1016/j.ipm.2023.103382.

[19] H. Shin, H. J. Lee, and S. Cho, "General-use unsupervised keyword extraction model for keyword analysis," *Expert Syst. Appl.*, vol. 233, no. Jan., pp. 120889–120896, 2023, doi: 10.1016/j.eswa.2023.120889.

[20] M. Zhang, X. Li, S. Yue, and L. Yang, "An empirical study of TextRank for keyword extraction," *IEEE Access*, vol. 8, no. Jan., pp. 178849–178858, 2020, doi: 10.1109/ACCESS.2020.3027567.

[21] C. Florescu and C. Caragea, "Positionrank: an unsupervised approach to keyphrase extraction from scholarly documents," *Assoc. Comput. Linguist.,* vol. 2017, no. Jan., pp. 1105–1115, 2017, doi: 10.18653/v1/P17-1102.

[22] O. Alqaryouti, H. Khwileh, T. Farouk, A. Nabhan, and K. Shaalan, "Graph-based keyword extraction," *Intell. Nat. Lang. Process*., vol. 2018, no. Jan., pp. 159–172, 2018, doi: 10.1007/978-3-319-67056-0_9.

[23] J. Sawicki, M. Ganzha, and M. Paprzycki, "The state of the art of natural language processing—a systematic automated review of NLP literature using NLP techniques," *Data Intell.,* vol. 5, no. Mar., pp. 707–749, 2023, doi: 10.1162/dint_a_00213.

[24] K. Kanclerz, "What if ground truth is subjective? personalized deep neural hate speech detection," *Proc. NLP Workshop,* vol. 2022, no. Jan., pp. 37–45, 2022.

[25] C. P. Chai, "Comparison of text preprocessing methods," *Nat. Lang. Eng.*, vol. 29, no. Mar., pp. 509–553, 2023, doi: 10.1017/S1351324922000213.

[26] Y. HaCohen-Kerner, D. Miller, and Y. Yigal, "The influence of preprocessing on text classification using a bag-of-words representation," *PLoS One,* vol. 15, no. May, pp. e0232525–e0232535, 2020, doi: 10.1371/journal.pone.0232525.

[27] P. M. Rahate and M. Chandak, "An experimental technique on text normalization and its role in speech synthesis," *Int. J. Innov. Technol. Explor. Eng.,* vol. 8, no. Aug., pp. 545–548, 2019.

[28] D. J. Ladani and N. P. Desai, "Stopword identification and removal techniques on TC and IR applications: a survey," *Int. J. Adv. Comput. Commun. Sys*t., vol. 2020, no. Mar., pp. 466–472, 2020, doi: 10.1109/ICACCS48705.2020.9074166.

[29] A. D. Latief, T. Sampurno, and A. O. Arisha, "Next sentence prediction: the impact of preprocessing techniques in deep learning," *Int. J. Comput. Control Inform.,* vol. 2023, no. Jan., pp. 274–278, 2023, doi: 10.1109/IC3INA60834.2023.10285805.

[30] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, and C. D. Manning, "Stanza: a Python natural language processing toolkit for many human languages," *arXiv, v*ol. 2020, no. Mar., pp. 1–12, 2020, doi: 10.48550/arXiv.2003.07082.

[31] J. Petrus, Ermatita, Sukemi, and Erwin, "An adaptable sentence segmentation based on Indonesian rules," *IAES Int. J. Artif. Intell.,* vol. 12, no. Sep., pp. 1491–1499, 2023, doi: 10.11591/ijai.v12.i3.pp1491-1499.

[32] M. A. Rosid, A. S. Fitrani, I. R. I. Astutik, N. I. Mulloh, and H. A. Gozali, "Improving text preprocessing for student complaint document classification using sastrawi," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 874, no. Jan., pp. 12017–12024, 2020, doi: 10.1088/1757-899X/874/1/012017.

[33] S. Senbel, "Fast and memory-efficient TFIDF calculation for text analysis of large datasets," *Lect. Notes Comput. Sci*., vol. 12798, no. Jan., pp. 557–563, 2021, doi: 10.1007/978-3-030-79457-6_47.

[34] L. Afuan, N. Hidayat, N. Nofiyati, and M. F. As'ad, "Sentiment analysis of the Kampus Merdeka program on Twitter using support vector machine and a feature extraction comparison: TF-IDF vs. FastText," *J. Appl. Data Sci.,* vol. 5, no. Apr., pp. 1738–1753, 2024, doi: 10.47738/jads.v5i4.436.

[35] C. D. Manning, P. Raghavan, and H. Schütze, "Introduction to information retrieval," *Cambridge Univ. Press*, vol. 2008, no. Jan., pp. 1–500, 2008.

[36] H. N. Irmanda and S. Hartati, "Sentiment analysis of cyberbullying using machine learning," *Int. J. Inform. Multimedia Cyber Inf. Syst.*, vol. 2024, no. Jan., pp. 594–600, 2024, doi: 10.1109/ICIMCIS63449.2024.10957620.

[37] J. Shen, "Utilizing natural language processing for efficient text analysis in the era of social media," *IEEE China Conf. Syst. Simul. Technol.,* vol. 2024, no. Jan., pp. 320–325, 2024, doi: 10.1109/CCSSTA62096.2024.10691866.

[38] S. Akter, F. M. J. M. Shamrat, S. Chakraborty, A. Karim, and S. Azam, "COVID-19 detection using deep learning algorithm on chest X-ray images," *Biology,* vol. 10, no. Nov., pp. 1–12, 2021, doi: 10.3390/biology10111174.

[39] P. Mishra, C. M. Pandey, U. Singh, A. Gupta, C. Sahu, and A. Keshri, "Descriptive statistics and normality tests for statistical data," *Ann. Card. Anaesth*., vol. 22, no. Jan., pp. 67–72, 2019, doi: 10.4103/aca.ACA_157_18.