Adjusted TextRank for keyword extraction in petrochemical project correspondence documents

Indri Atmoko, Evi Yulianti, Meganingrum Arista Jiwanggi
Faculty of Computer Science, Universitas Indonesia, Depok, Indonesia

Article Info

Article history:

Received Jan 22, 2024
Revised Mar 29, 2024
Accepted Apr 7, 2024

Keywords:

Bahasa Indonesia
Keyword extraction
Phrase detection
Project management
TextRank

ABSTRACT

A large petrochemical construction project is typically executed by multiple parties, all bound by contract agreement. During the execution phase, issues and problems may arise because the work details are not clearly specified in the contractual agreement. These issues are formally communicated and documented through written correspondence letters. By identifying important keywords within these formal letters, a comprehensive narrative of the project, including its associated issues, can be identified and analyzed. In this research, we introduce an adjusted TextRank algorithm that integrates external features from the Indonesian FastText language model and term frequency-inverse document frequency (TF-IDF) scores to identify important keywords within a dataset of correspondence letters of petrochemical projects. This enhancement involves refining phrase detection, semantic relationship estimation between words, and part-of-speech (POS) identification for words or phrases. Our results show that the proposed adjustments result in improved evaluation scores compared to the baseline standard TextRank and standard TF-IDF, respectively by 24.1% and 25% in terms of F-1 scores.

Corresponding Author:

Evi Yulianti
Faculty of Computer Science, Universitas Indonesia
Depok, Indonesia
Email: evi.y@cs.ui.ac.id

1. INTRODUCTION

In petrochemical industries, project execution requires collaboration between multiple parties bound by contractual agreements. However, during the implementation phase, unforeseen challenges often emerge that are not directly covered by the contract, requiring in-depth analysis for resolution. These issues are formally communicated and documented throughout the duration of the project, typically in formal written correspondence. The most common form of formal written communication is a letter, which generally raises an issue or provides a response to a problem.

Petrochemical projects can involve a considerable number of correspondence letters, potentially reaching thousands of letters. Extracting important keywords from these letters can benefit the project stakeholders because it can provide a comprehensive understanding of the historical project's issues [1]. By accurately extracting important keywords from project formal communication, we gain deeper insights into potential bottlenecks and can make more accurate decisions for the next step in project execution. This will greatly assist the project management team in prioritizing problem-solving and creating an execution plan for the project. Within Indonesian projects, especially in the petrochemical sector, keyword extraction techniques for mapping project challenges are underutilized. This is likely due to a lack of established practices and research. This study addresses this gap by applying keyword extraction techniques to a collection of Indonesian-language correspondence from petrochemical projects.

The challenge is that keyword extraction frequently relies on supervised training with English annotated datasets. This presents a significant obstacle for Indonesian-language documents, as there's a limited availability of annotated Indonesian datasets–especially within the petrochemical domain. Therefore, utilizing unsupervised techniques, which do not depend on annotated datasets, becomes a suitable approach. In this study, we propose a keyword extraction method based on TextRank with enhancements that can improve performance for Indonesian documents in the context of project management in petrochemical projects.

TextRank is a common method that has been used in previous work for keyword extraction. TextRank is a graph-based algorithm that estimates the importance of keywords based on its relationship with all keywords in a document. Previous research demonstrates the adaptability of the TextRank algorithm for keyword extraction across various languages and domains [2]-[4] specifically focused on Mandarin Chinese and a transportation industry dataset, incorporating word embeddings and weighting factors based on document features. Mao *et al.* [5] developed English text abstraction, enhancing TextRank with word co-occurrence features, semantic relationships from WordNet, word embeddings, and normalized Google Distance. Other notable research includes [6] combining term frequency-inverse document frequency (TF-IDF) and TextRank with Mandarin news articles.

Additional research includes [7] explored pointwise mutual information weighting on an English dataset from the NetEase website to calculate statistical correlations between parameters. Zhou *et al.* [8] used a Wikipedia Chinese corpus with a TextRank method based on rough data-deduction. Miah *et al.* [9] working with science papers in Portuguese, English, and Spanish, employed TextRank with keyphrase concentrated area (KCA) identification. Dhanasekar [10] applied TextRank to Arabic academic articles across disciplines, leveraging graph direction, weighted vertices using word co-occurrences and other document features. Wongchaisuwat [11] employed TextRank with combined sentence and word importance scores on English science project research abstracts. Xu and Zhang [12] fused TF-IDF and TextRank with a Mandarin news dataset, while [13] applied a similar fusion with modified TF-IDF weighted by long short-term memory (LSTM) classification on a Chinese news library dataset.

These research highlight limitations within the standard TextRank algorithm. Specifically, word relationships are often restricted to a single document, missing broader semantic connections. Additionally, the algorithm may overlook the importance of part-of-speech (POS) tags in determining word relevance and underestimate the significance of phrases. Despite these limitations, TextRank demonstrates potential for adjustments and can be combined with other techniques to enhance its performance within specific language and domain contexts.

Methods for improving the relationships between words in graph models involve adjusting vertices. These adjustments may include utilizing Word2Vec embedding similarities [14], the Jaccard similarity coefficient [15], and modifying word co-occurrence windows [16]. This study investigates the effects of combining word co-occurrences with word embeddings. While earlier research has explored the impact of single vertex weighting models, they have not explicitly addressed the influence of combining factors that can affect the connections between keyword candidates. This study combines word co-occurrences to obtain internal document proximity values and leverages word embedding similarities to calculate external word proximity within the *Bahasa Indonesia* domain.

To improve semantic connections, researchers have successfully used word embeddings, pretrained on larger corpora within the same language domain [17]-[22]. This study proposes a unique approach: utilizing word embeddings trained on a large *Bahasa Indonesia* corpus and employing FastText to minimize the occurrence of unknown vocabulary words, which can be prevalent in specialized domains like petrochemical projects. The selection of FastText embedding for this study is due to its ability to leverage subword information, namely n-grams, which significantly enhances its capability to manage out-of-vocabulary (OOV) words [23]. This feature is particularly valuable for domain-specific datasets, like those in the petrochemical project domain addressed in this study, which often contain specific vocabularies that are not commonly found in general text documents (i.e., Wikipedia) used to pretrain the FastText models. FastText's approach to generating embeddings for OOV words based on their constituent n-grams makes it an ideal choice for handling the unique and less common vocabulary encountered in specialized fields such as petrochemical projects.

Additionally, the traditional TextRank method frequently overlooks the importance of phrases. Phrase detection plays a crucial role as keywords often appear within phrases, carrying more meaning than individual words. Sun *et al.* [18] explored an approach for Mandarin texts by analyzing frequently recurring word groups in a document. Gunawan *et al.* [24] applied a method to Indonesian texts using the Figuera tool to synthesize phrases after computing TextRank scores. Within the Thai language domain [25], successfully implemented POS-based phrase detection. Unlike previous research, this study uses phrase detection trained for *Bahasa Indonesia* based on POS patterns. This means that phrases are identified directly within the document, ensuring no external or synthesized phrases are included. This method yields more common phrases compared to frequency-based approaches.

POS tagging plays a crucial role, as nouns and verbs tend to be more significant keywords compared to other parts of speech [26]-[28]. This study implements a similar technique, utilizing a tool specifically trained for *Bahasa Indonesia* to ensure accuracy. Additionally, this study implements a weighting system based on keyword position, assigning different levels of importance depending on where the word is located within the text [29].

Further, we also combine the proposed adjusted TextRank method with TF-IDF by weighting the keyword scores given by the adjusted TextRank method with its TF-IDF scores. We argue that this TF-IDF score integration may help to penalize common words/phrases, which appear very frequently in the entire corpus, that are potentially not important keywords, and give more incentives to the more important keywords [30]-[33]. Overall, this research aims to answer the following questions: 1) How does the performance of adjusted TextRank, compared to standard TF-IDF and standard TextRank for extracting keywords from Indonesian language documents specifically in the petrochemical project domain? 2) How can we create a fusion between TF-IDF and adjusted TextRank to further improve the performance of the adjusted TextRank?

This research is expected to contribute to the field of construction project management practice by enhancing the keyword extraction process from project correspondence documents. It aims to identify the important issues raised in the documents to support the project execution to proceed smoothly, on time, within budget, and benefit all involved parties. Extracting keywords that represent real project correspondence letters will complement the analysis of topics in the application of natural language processing (NLP) in construction project management. This research addresses practical challenges, including accurately detecting Indonesian phrases and combining graph-based weight analysis with semantic features from pre-trained word embeddings. It introduces a novel adjusted TextRank approach, demonstrating the effectiveness of combining it with TF-IDF for keyword extraction.

2. RESEARCH METHOD

The research method follows a four-stage process: data collection, preprocessing, experimentation, and evaluation. During data collection, PDF letter documents are gathered and annotated. The dataset is then cleaned and formatted for experiments in the preprocessing stage. Next, baseline experiments and proposed methods are evaluated. Finally, the outcomes are analyzed and conclusions are formed during the evaluation stage.

2.1. Dataset collection

The dataset used in this study is the collection of correspondence letters of construction projects from a petrochemical company in Indonesia. It contains the correspondence letters from the beginning of a project up to its near completion. The dataset is in the form of standard project PDF files, complete with logos, letter identification numbers, titles, and content. The content of the letters is in *Bahasa Indonesia*. This PDF dataset was then extracted and transformed into a text format to facilitate further processing. The extracted text are the letter numbers, titles, and contents.

For our experiment, a total of 1,000 letters were selected. Keywords annotation was then performed to produce ground truth keywords from each letter in our dataset. The results of the manual annotations served as the gold standard for evaluating the effectiveness of our keyword extraction methods. The annotation team consisted of two individuals with petrochemical industry experience. Therefore, they are already familiar with reading the project correspondence letters and they can easily identify the important keywords highlighted in the letters. Each annotator performs the annotation to a total of 550 letters, with an overlap of 50 letters. The overlap annotation was conducted to compute the agreement between annotators. To perform the annotation process, the annotators need to thoroughly read each letter in its entirety to accurately identify important keywords.

The statistics of our exploration results of the dataset are tabulated in the Table 1. We can see that on average, our documents are short or medium in length, with the average length is 9.6 sentences. This matches a typical length of the correspondence letters. Then, on average, there are 3.6 important keywords contained in each letter. Cohen's Kappa analysis was performed to calculate the agreement score between the two independent annotators. The resulting score of 0.51 indicates a moderate level of agreement.

Table 1. Statistical exploration of the dataset

Parameter	Statistics
Total documents	1,000 documents
Average number of sentences per document	9.6 sentences
Average number of words per document	210 words
Gold truth	1-7 words Average of 3.6 keywords per document

2.2. Pre-processing

The purpose of the preprocessing method is to obtain a clean dataset and format the dataframe so it can be used as input for our keyword extraction method. The steps of preprocessing are illustrated in Figure 1. The tasks carried out in this phase include text cleaning for symbols, numbers, and punctuation characters, converting all fonts to lowercase, and word tokenization. The tokenization involves breaking down sentences into their basic elements such as words and phrases, without changing the basic form of the words through stemming.

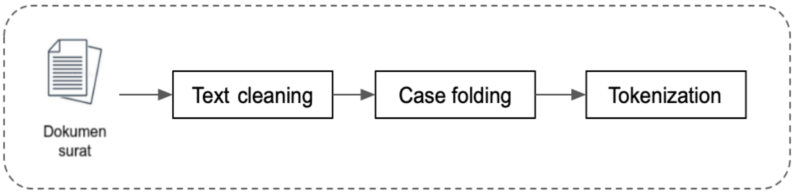

Figure 1. Preprocessing pipeline for the experiment

2.3. Keyword extraction methods

The proposed algorithm is constructed as a pipeline based on a variation of the TextRank model. Generally, traditional TextRank is a ranking algorithm for words, modeled in the form of a graph. TextRank is notably flexible, as it can be applied to various languages without altering its fundamental algorithm, owing to its independence from training data for document processing. The formulation for calculating TextRank is modeled in (1).

$$WS(V_i) = (1-d) + d * \sum_{V_j \varepsilon Adj(V_j)} \frac{w_{ji}}{\sum_{V_k Adj(V_j)} w_{jk}} WS(V_j) \quad (1)$$

Where V_i is the vertex, whose score is being calculated, V_j is a vertex connected to V_i, V_k is a vertex connected to V_j. W_{ji} denotes the edge weight between vertices V_i and V_j, while W_{jk} represents the edge weight between vertices V_j and V_k. The term 'd' is the damping factor, which varies between 0 and 1, typically defaulting to a value of 0.85.

Two methods are proposed in this work: adjusted TextRank and adjusted TextRank+TF-IDF. The first method, i.e., adjusted TextRank, is expected to enhance the original TextRank method by refining the phrase detection, the understanding of semantic relationships between words, and the identification of POS tags for words or phrases. Next, the second method is expected to enhance our first method by integrating TF-IDF scores to the adjusted TextRank. More on the first and second methods are explained in the following subsections.

2.3.1. Adjusted TextRank

This method is proposed to address the first research question. The adjusted TextRank method is composed of a series of steps: detecting Indonesian phrases, representing these phrases in word embedding vectors, filtering keywords, and implementing weighting factors. Figure 2 illustrates the flow diagram of the proposed adjusted TextRank method which contains five new components that we add to the original TextRank method, such as POS tagging for phrase detection (Component-1), semantic similarity computation between keywords using word embedding (Component-2), POS tag filtering (Component-3), weighting of keywords based on position (Component-4), and weighting of phrases (Component-5). POS tags are acquired using a tool from nlp-id (https://github.com/kumparan/nlp-id).

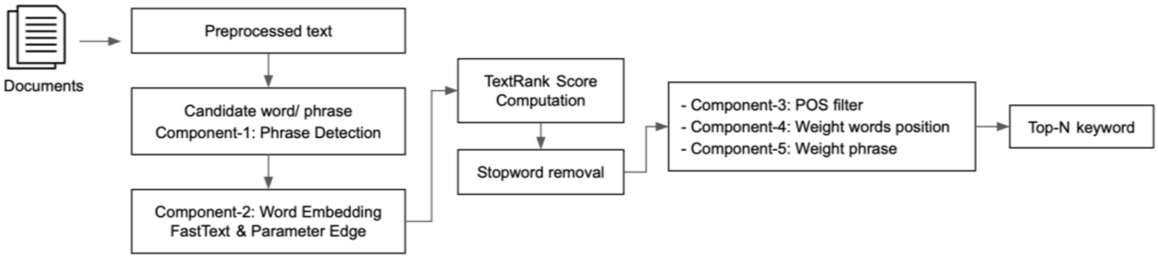
Documents
Preprocessed text
Candidate word/ phrase
Component-1: Phrase Detection
Component-2: Word Embedding
FastText & Parameter Edge
TextRank Score
Computation
Stopword removal
- Component-3: POS filter
- Component-4: Weight words position
- Component-5: Weight phrase
Top-N keyword

Figure 2. Adjusted TextRank pipeline

The general processing flow carried out by the adjusted TextRank method is as follows:

i) After an input document is preprocessed, then our system detects the phrases contained in the document using POS tagging techniques. Bigram (two-word) and trigram (three-word) combinations are considered as phrases in Indonesian if they meet the predetermined POS patterns. These phrases together with the unigram words serve as keywords that will be further processed by our system for keyword extraction tasks. POS tagging is performed using the nlp-id pretrained model, which is trained with common formation patterns for the Indonesian phrase, as detailed in (2).

$$\begin{aligned}
&DP = \{<NUM><NNP><NUM>\} \\
&NP \quad = \{<NNP><NNP>+\}/\{<NN>+<JJ>\}/\{<FW>+\}\{<NP><NP>+\} \\
&ADJP = \{<JJ><ADV>\} \,/\, \{<ADV><JJ>\} \,/\, \{<JJ>+\} \,/\, \{<NEG>*<ADJP>\} \\
&NUMP = \{<NUM><NUM>+\} \qquad (2)
\end{aligned}$$

where DP (date phrase), NUM (number), NNP (proper noun), NN (noun), JJ (adjective), FW (foreign word), ADJP (adjective phrase), ADV (adverb), NEG (negation), VP (verb phrase), and VB (verb).

ii) After words and phrases are detected, their embeddings (vector representations) are generated using a pretrained Indonesian FastText model. These embeddings enable the calculation of semantic relationships among keywords. Keyword proximity is calculated using cosine similarity values.

$$cos\, cos\, (\emptyset) = \frac{A*B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} * \sqrt{\sum_{i=1}^{n} B_i^2}} \qquad (3)$$

where A_i and B_i are the ith components of vectors A and B, respectively.
The weights of the edges in the graph are obtained by multiplying the cosine similarity between two keywords by their co-occurrence count.

$$W_{ij} = cos\, (\emptyset)_{ij} * P(w_i, w_j) \qquad (4)$$

where $P(w_i, w_j)$ is the co-occurence rate of word w_i and w_j.
This weighting schema enables the graph model to identify not only frequently co-occurring words/phrases but also to capture the semantic closeness between them. The FastText embedding model to be used in this study has been trained using a dataset from Indonesian Wikipedia with a vector size of 200. The reason for choosing FastText methods is its ability to tackle the issue of OOV problem by estimating the vector representation of unseen words by taking the total of vector representation of its subwords (n-gram characters). This greatly reduces problems with words that are not in the Wikipedia vocabularies, which is especially beneficial for datasets in specific domains. The petrochemical project documents may contain some specific words that rarely occur in the general Wikipedia collection.

iii) A graph representation of keywords is constructed and the importance scores for each keyword are computed using the TextRank algorithm formula described in section 2.3.

iv) Stopwords are filtered out using a stopwords obtained from [34] and Sastrawi (https://github.com/sastrawi/sastrawi). The filtering applies only to unigrams. Standalone stopwords, which usually lack significant meaning, are removed during filtering. However, stopwords within meaningful phrases are retained to preserve context.

v) A specific POS filtering technique is designed to prioritize candidate words or phrases that are nouns, verbs, or foreign words. This is achieved by focusing on POS tags such as NN, NNP, VB, NP (noun phrase), VP, and FW. Words or phrases that do not belong to this POS are removed.

vi) Keyword importance scores are weighted based on word positions. Because the title of a document often carries important meaning, then we assign extra weight to the importance scores of words or phrases that are found in the title section of a document. Title is given a weighting factor of 2 in the score calculation.

vii)The importance scores of phrases are weighted higher because they often convey more meaning and can more effectively represent the main topic or issue in a text. In the TextRank method, phrases tend to obtain lower TextRank scores than individual words because they usually interact less frequently with other words in word co-occurrence values. Therefore, in this step, we provide a scoring incentive for phrases. The weight factor assigned to phrases is 2 for bigrams and 3 for trigrams.

2.3.2. Combination of adjusted TextRank and TF-IDF

This method is proposed to address the second research question. In this method, we add a new component (Component-6) that combines the adjusted TextRank scores and TF-IDF scores. This integration is conducted for each candidate word or phrase within a single document. This approach allows the combined score to reflect a word's importance, both in terms of its semantic structure (TextRank) and its frequency of occurrence (TF-IDF). Here, the TextRank scores are computed using the adjusted TextRank method described earlier in section 2.3.1. Then, the resulting score for each keyword is then weighted by the TF-IDF score of that keyword using a simple multiplication.

$$Score(w_i) = ws(v_i) * tfidf_i \tag{5}$$

where $Score(w_i)$ is word score for combination of adjusted TextRank and TF-IDF, $ws(v_i)$ is word score adjusted TextRank and $tfidf_i$ is word score calculating using TF-IDF.

2.4. Evaluation metric

The evaluation of keyword extraction is generally measured by accuracy value P (precision), recall value R (Recall) and average value F1. We specially use F1 as the main criteria due to the F1 as the micro average of P and R, can reflect the results more accurately. The calculation formula is as(6) to (8).

$$P = \frac{\lfloor EK \cap MANU \rfloor}{\lfloor EK \rfloor} \tag{6}$$

$$R = \frac{\lfloor EK \cap MANU \rfloor}{\lfloor MANU \rfloor} \tag{7}$$

$$F1 = \frac{2PR}{P+R} \tag{8}$$

In this context, EK is the extracted keyword from the conducted experiments, and MANU represents the manually derived keyword that serves as the gold standard reference.

3. RESULTS AND DISCUSSION

The experiment results are presented in a structured experimental scenario outlined below. Several scenarios were repeated to set them as the baseline for subsequent scenario experiments. The experiment result is displayed in both tabular forms and analyzed comprehensively as follows.

3.1. The performance of adjusted TextRank

We set our baseline experiment using the basic versions of TF-IDF and TextRank, without any enhancement techniques for a *Bahasa Indonesia* dataset. In this experiment, we used a basic computational approach to identify phrases in a sentence. Bigrams and trigrams that appear more than twice are identified as the valid phrase [16].

We then compare the baseline results with our proposed method, the adjusted TextRank. We found that standard TF-IDF achieves a slightly better F1 score compared to TextRank for this dataset. A weakness of the standard TextRank is its focus on a single document, whereas TF-IDF benefits from the IDF factor that considers the entire corpus. The enhancement in our proposed method includes phrase detection, improving graph edge modeling, applying POS filtering, and implementing a special node enhancement score. The special node is designated for candidate keywords, based on position, phrase type, or as a single word. Through these enhancements, there was an increase in the F1 score, rising from 0.250 to 0.287. Comparative results versus baseline are tabulated in Table 2. Our results demonstrate that these adjustments improve keyword extraction performance.

Table 2. Adjusted TextRank experiment result

Method	Precision	Recall	F1 score
TF-IDF	0.416	0.179	0.250
TextRank	0.402	0.172	0.241
Adjusted TextRank	**0.478**	**0.205**	**0.287**

3.2. The performance of adjusted TextRank+TF-IDF

The objective of the second system experiment was to determine whether the TF-IDF score could be effectively used as a weighting factor for the previously proposed adjusted TextRank score, addressing the second research question of this study. The results indicate that the introduction of an additional feature, which computes the TF-IDF and subsequently integrates it with the modified TextRank from the initial system experiment, resulted in improved performance in keyword extraction. This improvement elevated the score from 0.287 to 0.300.

Our study suggests that the TF-IDF weight factor compensates for the weaknesses of the current adjusted TextRank method. Both the baseline TextRank and the adjusted version use vertex representation based on word co-occurrence in a single document. This means that words appearing frequently in the document are likely to receive a higher score. However, this approach does not take the overall document context into consideration. TF-IDF, which considers the context of the entire document compensates for this limitation. The detailed comparative evaluation can be found in Table 3.

Table 3. Adjusted TextRank+TF-IDF experiment result

Method	Precision	Recall	F1 score
Adjusted TextRank	0.478	0.205	0.287
Adjusted TextRank+TF-IDF	**0.499**	**0.214**	**0.300**

3.3. Ablation analysis

An ablation analysis is a technique that involves removing components of a system to understand how each component contributes to the system's overall performance. We conducted an ablation analysis on our best-performing method to analyze its components. This study analyzed the importance of each new component added to the original TextRank method. While earlier research has explored the impact of overall combined enhancements, they have not explicitly addressed the influence of each individual component. We remove or "ablate" each individual component to observe how their removal affects overall performance. In this study, we will observe the importance of each of six new components that we add to the Adjusted TextRank+TF-IDF method.

In our ablation study, we used the adjusted TextRank combined with TF-IDF as the baseline, with results presented in Table 4. We sequentially removed components and observed the following performance changes: Component-6 (+0.013), Component-5 (+0.056), Component-4 (+0.010), Component-3 (+0.001), Component-2 (+0.007), and Component-1 (+0.059). Our analysis indicates that phrase identification (Component-1) and phrase weighting (Component-5) are the primary drivers behind the improved keyword extraction results. Each of these components contributes uniquely to improving the accuracy and detail of our results. Gagliardi and Artese [17] suggestion, we generated keyphrase candidates before ranking, resulting in improved performance by identifying phrases directly from the document rather than synthesizing them.

3.4. Qualitative analysis

Through this qualitative study, we would like to highlight various aspects of keyword extraction methods, in terms of phrase detection quality, contextual and topic relevance and suitability for niche domains like petrochemicals construction. For illustrative comparison examples, please see Table 5. Adjusted TextRank excels in phrase detection, identifying more common *Bahasa Indonesia* phrases and context-specific phrases such as "*perangkat lunak*" (software) and "payment milestone,". This contrasts with the TF-IDF method, which sometimes yields less relevant or awkward phrases such as ‘*pemecahan payment*’, ‘*penggantian oleh*’, and ‘*perusahaan biaya*’. Meanwhile, basic TextRank often selects top-n keywords in unigram form due to its reliance on co-occurrences, where unigrams naturally have a higher probability than n-grams.

When considering context and topic relevance, adjusted TextRank+TF-IDF aligns more closely with the gold standard. For instance, in topics like software management tools for system completion, both adjusted TextRank and adjusted TextRank+TF-IDF predict keywords that are relevant to this domain. Similarly, for subjects like international logistics supply chain and financial issues, these methods again show greater relevance to the topics.

Recent experiments suggest that the combined method of adjusted TextRank and TF-IDF shows promise in keyword extraction within specific domains. Our findings suggest that this technique effectively balances precise phrase detection and topic relevance, without solely relying on term frequency. Further improvements could involve using modified TF-IDF with hierarchical methods [31] or weighting scheme [16].

Table 4. Ablation study experiment result

Ablation component	Precision	Recall	F1 score	ΔF1
Adjusted TextRank+TFIDF	0.499	0.214	0.300	0.000
Component-6	0.478	0.205	0.287	0.013
Component-5	0.406	0.174	0.244	0.056
Component-4	0.484	0.207	0.290	0.010
Component-3	0.498	0.214	0.299	0.001
Component-2	0.488	0.209	0.293	0.007
Component-1	0.401	0.172	0.241	**0.0059**

Table 5. The example of keyword extraction results

Goldtruth	TF-IDF	TextRank	Adjusted TextRank	Adjusted TextRank+TF-IDF
System completion; *Jawaban*; *Penggunaan*; Proxis	*Perangkat*; Contractor; **Proxis**	Utilization; **Completion**; Management	*Perangkat*; *Perangkat lunak;* **System**	**System completion**; *Perangkat;* *Perangkat lunak*

Control valve;
Permohonan;
Material eksotis;
Izin ekspor
	Mto supplier;
	End user;
	Globe and
		Request;
		Signature;
		End user
			User statement;
			End;
			Check valve
				End;
				Control valve supplier;
				Supplier

Kendala finansial;
Payment milestones;
Side letter;
Biaya;
Pemecahan payment
Milestones
	Belum dilakukan;
	Penggantian oleh;
	Perusahaan biaya
		Ke empat;
		Usulan perubahan;
		Atas payment
			Perubahan pekerjaan baik;
			Payment milestone;
			Penandatanganan;
			Perubahan ketiga
				Pembayaran;
				Biaya;
				Payment milestone

4. CONCLUSION

Letters hold valuable insights into project narratives and challenges within the petrochemical construction sector. Due to their lack of language-specific annotated datasets and specialized non-mainstream topics, unsupervised algorithms are particularly suitable for keyword extraction from *Bahasa Indonesia* letters. While statistical methods like TF-IDF initially perform slightly better, TextRank’s flexibility offers potential for improvement. This study demonstrates that enhancements to TextRank's node/edge weights and context integration can outperform the TF-IDF baseline by 3.7%. Combining it with TF-IDF yields an even greater 5% performance boost. Ablation analysis pinpoints phrase detection as the most significant contributor to this improvement, with a shift from frequency-based to POS pattern-based methods offering the most substantial score increase. These findings highlight the promise of graph-based keyword extraction for niche domains and languages, suggesting that refined phrase detection and representation techniques can further enhance effectiveness and accuracy.

ACKNOWLEDGEMENTS

This research was funded by the Directorate of Research and Development, Universitas Indonesia, under Hibah PUTI Pascasarjana 2024 (Grant No. NKB-28/UN2.RST/HKP.05.00/2024).

REFERENCES

[1] H. Yan, N. Yang, Y. Peng, and Y. Ren, “Data mining in the construction industry: present status, opportunities, and future trends,” *Automation in Construction*, vol. 119, p. 103331, Nov. 2020, doi: 10.1016/j.autcon.2020.103331.

[2] M. Garg, “A survey on different dimensions for graphical keyword extraction techniques,” *Artificial Intelligence Review*, vol. 54, no. 6, pp. 4731–4770, Aug. 2021, doi: 10.1007/s10462-021-10010-6.

[3] E. Papagiannopoulou and G. Tsoumakas, “A review of keyphrase extraction,” *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 10, no. 2, 2020, doi: 10.1002/widm.1339.

[4] S. Zhang, R. Ye, J. Wang, and K. Yang, “A keyword extraction method for transportation industry standards based on improved TextRank,” in *2021 IEEE 3rd International Conference on Frontiers Technology of Information and Computer (ICFTIC)*, Nov. 2021, pp. 236–241, doi: 10.1109/ICFTIC54370.2021.9647206.

[5] X. Mao, S. Huang, R. Li, and L. Shen, "Automatic keywords extraction based on co-occurrence and semantic relationships between words," *IEEE Access*, vol. 8, pp. 117528–117538, 2020, doi: 10.1109/ACCESS.2020.3004628.

[6] L. Yao, Z. Pengzhou, and Z. Chi, "Research on news keyword extraction technology based on TF-IDF and TextRank," in *2019 IEEE/ACIS 18th International Conference on Computer and Information Science (ICIS)*, Jun. 2019, pp. 452–455, doi: 10.1109/ICIS46139.2019.8940293.

[7] Y. Tao, Z. Cui, and Z. Jiazhe, "Research on keyword extraction algorithm using PMI and TextRank," in *2019 IEEE 2nd International Conference on Information and Computer Technologies (ICICT)*, Mar. 2019, pp. 5–9, doi: 10.1109/INFOCT.2019.8711099.

[8] N. Zhou, W. Shi, R. Liang, and N. Zhong, "TextRank keyword extraction algorithm using word vector clustering based on rough data-deduction," *Computational Intelligence and Neuroscience*, vol. 2022, pp. 1–19, Jan. 2022, doi: 10.1155/2022/5649994.

[9] M. B. A. Miah, S. Awang, M. S. Azad, and M. M. Rahman, "Keyphrases concentrated area identification from academic articles as feature of keyphrase extraction: a new unsupervised approach," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 1, 2022, doi: 10.14569/IJACSA.2022.0130192.

[10] S. Dhanasekar, "Keyword extraction from Arabic text using the page rank algorithm," *International Journal of Innovative Technology and Exploring Engineering*, vol. 8, no. 12, pp. 3495–3504, Oct. 2019, doi: 10.35940/ijitee.L2614.1081219.

[11] P. Wongchaisuwat, "Automatic keyword extraction using TextRank," in *2019 IEEE 6th International Conference on Industrial Engineering and Applications (ICIEA)*, Apr. 2019, pp. 377–381, doi: 10.1109/IEA.2019.8714976.

[12] Z. Xu and J. Zhang, "Extracting keywords from texts based on word frequency and association features," *Procedia Computer Science*, vol. 187, pp. 77–82, 2021, doi: 10.1016/j.procs.2021.04.035.

[13] X. Ao, X. Yu, D. Liu, and H. Tian, "News keywords extraction algorithm based on TextRank and classified TF-IDF," in *2020 International Wireless Communications and Mobile Computing (IWCMC)*, Jun. 2020, pp. 1364–1369, doi: 10.1109/IWCMC48107.2020.9148491.

[14] S. Song, Z. Wang, S. Xu, S. Ni, and J. Xiao, "A novel text classification approach based on Word2vec and TextRank keyword extraction," in *2019 IEEE Fourth International Conference on Data Science in Cyberspace (DSC)*, Jun. 2019, pp. 536–543, doi: 10.1109/DSC.2019.00087.

[15] H. T. Huynh, N. Duong-Trung, D. Q. Truong, and H. X. Huynh, "Vietnamese Text classification with TextRank and Jaccard similarity coefficient," *Advances in Science, Technology and Engineering Systems Journal*, vol. 5, no. 6, pp. 363–369, 2020, doi: 10.25046/aj050644.

[16] M. Zhang, X. Li, S. Yue, and L. Yang, "An empirical study of TextRank for keyword extraction," *IEEE Access*, vol. 8, pp. 178849–178858, 2020, doi: 10.1109/ACCESS.2020.3027567.

[17] I. Gagliardi and M. T. Artese, "Semantic unsupervised automatic keyphrases extraction by integrating word embedding with clustering methods," *Multimodal Technologies and Interaction*, vol. 4, no. 2, p. 30, Jun. 2020, doi: 10.3390/mti4020030.

[18] J. Sun *et al.*, "Text visualization for construction document information management," *Automation in Construction*, vol. 111, p. 103048, Mar. 2020, doi: 10.1016/j.autcon.2019.103048.

[19] Z. Zhang, J. Petrak, and D. Maynard, "Adapted TextRank for term extraction: a generic method of improving automatic term extraction algorithms," *Procedia Computer Science*, vol. 137, pp. 102–108, 2018, doi: 10.1016/j.procs.2018.09.010.

[20] Y. Wen, H. Yuan, and P. Zhang, "Research on keyword extraction based on Word2Vec weighted TextRank," in *2016 2nd IEEE International Conference on Computer and Communications (ICCC)*, Oct. 2016, pp. 2109–2113, doi: 10.1109/CompComm.2016.7925072.

[21] H. Kuang, H. Chen, X. Ma, and X. Liu, "A keyword detection and context filtering method for document level relation extraction," *Applied Sciences*, vol. 12, no. 3, p. 1599, Feb. 2022, doi: 10.3390/app12031599.

[22] C. Dascalu and S. Trausan-Matu, "Experiments with Contextualized word embeddings for keyphrase extraction," in *2021 23rd International Conference on Control Systems and Computer Science (CSCS)*, May 2021, pp. 447–452, doi: 10.1109/CSCS52396.2021.00079.

[23] D. Mahata, J. Kuriakose, R. R. Shah, and R. Zimmermann, "Key2Vec: automatic ranked keyphrase extraction from scientific articles using phrase embeddings," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, 2018, pp. 634–639, doi: 10.18653/v1/N18-2100.

[24] D. Gunawan, F. Purnamasari, R. Ramadhiana, and R. F. Rahmat, "Keyword extraction from scientific articles in bahasa Indonesia using TextRank algorithm," in *2020 4rd International Conference on Electrical, Telecommunication and Computer Engineering (ELTICOM)*, Sep. 2020, pp. 260–264, doi: 10.1109/ELTICOM50775.2020.9230514.

[25] R. Kedtiwerasak, E. Adsawinnawanawa, P. Jirakunkanok, and R. Kongkachandra, "Thai keyword extraction using TextRank algorithm," in *2019 14th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)*, Oct. 2019, pp. 1–6, doi: 10.1109/iSAI-NLP48611.2019.9045523.

[26] H. Zhao and Q. Xie, "An improved TextRank multi-feature fusion algorithm for keyword extraction of educational resources," *Journal of Physics: Conference Series*, vol. 2078, no. 1, p. 012021, Nov. 2021, doi: 10.1088/1742-6596/2078/1/012021.

[27] K. Kurniawan and A. F. Aji, "Toward a standardized and more accurate indonesian part-of-speech tagging," in *2018 International Conference on Asian Language Processing (IALP)*, Nov. 2018, pp. 303–307, doi: 10.1109/IALP.2018.8629236.

[28] C. Xu, "Research on Information retrieval algorithm based on TextRank," in *2019 34rd Youth Academic Annual Conference of Chinese Association of Automation (YAC)*, Jun. 2019, pp. 180–183, doi: 10.1109/YAC.2019.8787615.

[29] C. Zhang, L. Zhao, M. Zhao, and Y. Zhang, "Enhancing keyphrase extraction from academic articles with their reference information," *Scientometrics*, vol. 127, no. 2, pp. 703–731, Feb. 2022, doi: 10.1007/s11192-021-04230-4.

[30] W. Wang, X. Li, and S. Yu, "Chinese Text keyword extraction based on Doc2vec and TextRank," in *2020 Chinese Control And Decision Conference (CCDC)*, Aug. 2020, pp. 369–373, doi: 10.1109/CCDC49329.2020.9164788.

[31] E. Yulianti, N. Pangestu, and M. A. Jiwanggi, "Enhanced TextRank using weighted word embedding for text summarization," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 13, no. 5, pp. 5472–5482, Oct. 2023, doi: 10.11591/ijece.v13i5.pp5472-5482.

[32] B. Chiraratanasopha, S. Boonbrahm, and T. Theeramunkong, "Effect of term weighting on keyword extraction in hierarchical category structure," *Computing and Informatics*, vol. 40, no. 1, pp. 57–82, 2021, doi: 10.31577/cai_2021_1_57.

[33] S. Anjali, N. M. Meera, and M. G. Thushara, "A graph based approach for keyword extraction from documents," in *2019 Second International Conference on Advanced Computational and Communication Paradigms (ICACCP)*, Feb. 2019, pp. 1–4, doi: 10.1109/ICACCP.2019.8882946.

[34] F. Tala, "A study of stemming effects on information retrieval in bahasa Indonesia," Master of Logic Project, Institute for Logic, Language and Computation, Universiteit van Amsterdam, The Netherlands, 2003.

BIOGRAPHIES OF AUTHORS

Indri Atmoko received the S.T. degree in Electrical Engineering from Institute Technology of Bandung in 2005. He is now furthering his academic pursuits by studying for his Master's degree in Computer Science at the University of Indonesia. Skilled in more than just engineering. He is deeply involved in research, especially in implementing natural language processing (NLP) technology for managing projects, applying advanced computations in electrical power systems, and renewable energy solutions. His work uniquely combines electrical engineering, computer science, and project management. For any professional inquiries or collaborations. He can be contacted at email: indriatmoko@gmail.com.

Evi Yulianti is a lecturer and researcher at the Faculty of Computer Science, Universitas Indonesia. She received the B.Comp.Sc. degree from the Universitas Indonesia in 2010, the dual M.Comp.Sc. degree from Universitas Indonesia and Royal Melbourne Institute of Technology University in 2013, and the Ph.D. degree from Royal Melbourne Institute of Technology University in 2018. Her research interests include information retrieval and natural language processing. She can be contacted at email: evi.y@cs.ui.ac.id.

Meganingrum Arista Jiwanggi is a lecturer and researcher at Faculty of Computer Science, Universitas Indonesia. She received her bachelor’s degree in Computer Science from the Universitas Indonesia in 2012 and a dual master of Computer Science degree from the Universitas Indonesia and Royal Melbourne Institute of Technology University in 2014. Her research interests include natural language processing and data science. She can be contacted at email: meganingrum@cs.ui.ac.id.