Feature-based approach and sequential pattern mining to enhance quality of Indonesian automatic text summarization

Dian Sa'adillah Maylawati[1,2], Yogan Jaya Kumar[1], Fauziah Kasmin[1]
[1]Centre for Advanced Computing Technology, Faculty of Information and Communication Technology,
Universiti Teknikal Malaysia Melaka, Melaka, Malaysia
[2]Department of Informatics, Faculty of Science and Technology, UIN Sunan Gunung Djati Bandung, Bandung, Indonesia

Article Info

Article history:

Received Sep 5, 2022
Revised Feb 5, 2023
Accepted Feb 12, 2023

Keywords:

Automatic text summarization
Feature-based approach
Indonesian language
Prefixspan
Sentence scoring
Sequential pattern mining

ABSTRACT

Indonesian automatic text summarization research is developed rapidly. The quality, especially readability aspect, of text summary can be reached if the meaning of the text can be maintained properly. Therefore, this research aims to enhance the quality of extractive Indonesian automatic text summarization with considering the quality of structured representation of text. This research uses sequential pattern mining (SPM) to produce This research use SPM to produce sequence of words (SoW) as structured text representation using PrefixSpan algorithm. Then, SPM is combined with feature-based approach using sentence scoring method to produce summary. The experiment result using IndoSum dataset shows that even though the combination of SPM and sentence scoring can increase the precision value of recall-oriented understudy for gisting evaluation (ROUGE)-1, ROUGE-2, and ROUGE-L, from 0.68 to 0.76, 0.54 to 0.69, and 0.51 to 0.72. Especially, combination of SPM and Sentence Scoring can enhance precision, recall, and f-measure of ROUGE-L that consider the order of word occurance in measurement. SPM increases ROUGE-L f-measure value of sentence scoring from 0.32 to 0.36. Moreover, combination of sentence scoring and SPM is better than SumBasic that used as feature-based approach in the previous Indonesian text summarization research.

Corresponding Author:

Dian Sa'adillah Maylawati
Centre for Advanced Computing Technology, Faculty of Information and Communication Technology
Universiti Teknikal Malaysia Melaka
Street of Hang Tuah Jaya, Melaka, 76100, Malaysia
Email: diansm@uinsgd.ac.id

1. INTRODUCTION

The development of automatic text summarizing techniques is accelerating. natural language processing includes automatic text summarization (NLP) [1]. The study of natural language processing, or NLP, is a subfield of linguistics, computer science, and artificial intelligence that focuses on how to build computers that can process and analyze massive amounts of natural language input [2]-[4]. In general, there are two forms of automatic text summarization: extractive and abstractive [5]. Extractive summarization creates a sequential summary based on the document's source and without changing the word structure of the sentences [6], [7]. The source document sentences make up the final summary. Extractive summaries or extracts are made by detecting key sentences in the source content and selecting them directly. While abstractive summarization results in a modified summary [8], for example paraphrase. As a result, the abstracted summary does not contain the same phrases or structure as the original document, but it nevertheless conveys the same message.

In today's world, there are numerous approaches for text summarizing. There are at least three techniques to automatically producing summary, including feature-based approach. The feature-based technique considers and calculates document components (such as words, sentences, phrases, and so on) throughout the automatic text summarizing process [9]. Sentence scoring [10], [11], SumBasic [12], and latent semantic analysis (LSA) [13]-[15] are feature-based approach algorithms that can be utilized for text summarization in the previous research. However, the production of a readable summary is a challenge in automatic text summarizing research [16]-[18]. It means that there isn't much of a disconnect between the summary result and the reader's comprehension. Because it ensures that the summary result is readable and understood, the readability of the summary result is a critical aspect in evaluating the effectiveness of automatic text summarization [19]-[21].

Therefore, to resolve the readability issue from the summary results, it can be started from preparing a quality text representation. Text representation that used must be able to maintain the meaning of the text data. The sequence of words (SoW) is one of widely used as text representation for text analytic, either for text mining, information retrieval, text similarity, even for text summarization (both for single and multi-document) which is proven to be able to maintain the meaning of the text well and even increase. However, most of them are used for English, while each language has its own peculiarities, including the Indonesian language. Moreover, there is no specific research has been found using a SoW as a text representation for Indonesian automatic text summarization.

In previous research, the SoW used for the Indonesian language is not specific for text summarization [22]. While the SoW as text representation that maintains the meaning of the text and minimize the reader's understanding gap with the summary results, to reach readability of summary result. There is a text summarization study that uses the SoW for the Malaysian language [23], [24], which is one family with the Indonesian language, but there are still many differences. Therefore, this study aims to investigate and enhance the quality of Indonesian text summary result with SoW as text representation to maintain the meaning of the text. This research use PrefixSpan algorithm as one of sequential pattern mining (SPM) technique [25] to produce SoW and Sentence Scoring as feature-based approach to produce Indonesian automatic text summary. Then, the summary result will be evaluated with recall-oriented understudy for gisting evaluation (ROUGE-1), ROUGE-2, and ROUGE-L that many used in current Indonesian automatic text summarization [26]-[32]. This metrics is used to evaluate the quality of Indonesian automatic text summary that produced by the combination of SPM and sentence scoring.

2. MATERIALS AND METHOD

Figure 1 presents the research overview that combine SPM and sentence scoring to produce Indonesian automatic text summary. Combination of SPM in sentence scoring aims to enhance the performance of summary result and reach the readable summary. This research has several activities, begin from data gathering and data preparation, producing Indonesian automatic text summary using sentence scoring, producing Indonesian automatic text summary using combination of SPM and sentence scoring, evaluate the ROUGE value of summary result from sentence scoring with and without SPM, then comparative analysis whether SPM can enhance the quality of Indonesian automatic text summary result.

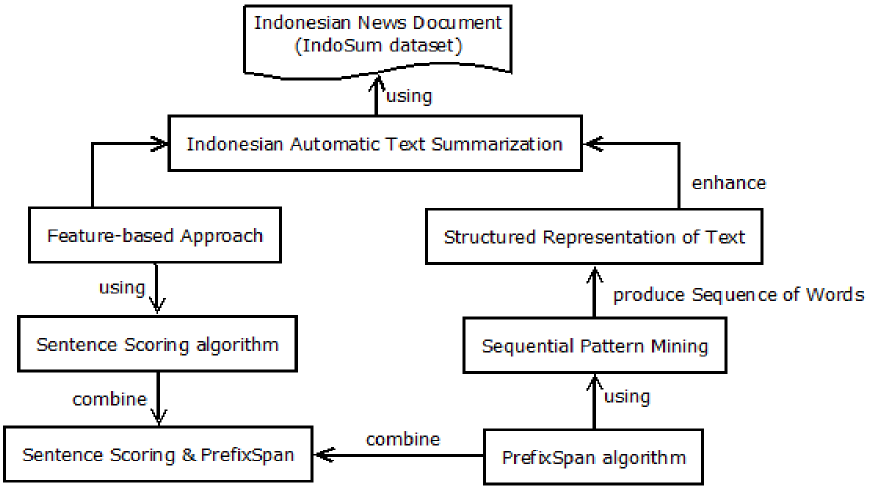

Indonesian News Document (IndoSum dataset)
using
Indonesian Automatic Text Summarization
enhance
Feature-based Approach
Structured Representation of Text
using
produce Sequence of Words
Sentence Scoring algorithm
Sequential Pattern Mining
combine
using
Sentence Scoring & PrefixSpan
combine
PrefixSpan algorithm

Figure 1. Research overview

Most of the related studies about Indonesian text summarization used news articles for an experiment because easily collected freely and online. Besides that, the most important thing is that news articles have a pretty good language structure or writing structure, compared to social media, which contains many languages with poor structure and even slang. This research uses IndoSum dataset that contain news articles from CNN Indonesia, Kumparan, and Merdeka.com [26]. This dataset is widely used as a benchmark for Indonesian automatic text summarization research. And also, the dataset have manual summary which created by two Indonesian native speakers as an expert. The IndoSum dataset is open access and available at Github [33]. The news articles dataset are classified by 6 categories: entertainment, inspiration, sport, showbiz, headline, and technology. Actually, from that sources total news article are 18,774 news documents. Indonesian news document collections are presented with JavaScript object notation (JSON) type.

The next process is data preparation or data pre-processing that significant steps in most computational linguistics investigations [34]. Raw texts should be pre-processed to ensure that they have a suitable representation and can be used effectively in experiments. There are several text preprocessing procedures in this study: separating the sentences, case-folding, tokenizing, removing regular expression (non-letter characters), removing Indonesian topwords, and stemming process using Nazief-Adriani algorithm for Indonesian language. In addition, there is Sastrawi library on Python programming that can be used to prepare Indonesian text data in the text pre-processing phase [35]. Sastrawi provides the Indonesian stop words list and Nazief-Adriani as a popular stemming process for Indonesian text [35], [36]. However, sometimes text pre-processing activity is not the same for all cases, depending on the need in each case. Sentence separation aims to distinguish one sentence from another according to the needs of the feature approach using sentence scoring.

Then, there are two experiment scenarios: i) produce Indonesian text summary with feature-based approach using sentence scoring algorithm and bag-of-words (BoW) as structured representation of text; and ii) produce Indonesian text summary with combinantion of feature-based approach using sentence scoring and PrefixSpan algorithm as a part of SPM, and SoW as structured representation of text.

The performance of SPM combination with feature-based is evaluated using co-selection-based analysis with ROUGE-1, ROUGE-2, and ROUGE-L. ROUGE also evaluate the precision, recall, and F-1 score. ROUGE-1 and ROUGE-2 is a part of ROUGE-N, where N=1, 2, 3, and 4. ROUGE-N score evaluation use n-gram overlaps between the candidate document and the reference documents. ROUGE-N formula is available in (1) [37].

$$ROUGE - N = \frac{\sum_{S\in\{Reference\ Summaries\}}\sum_{gram_n\in S} Count_{match}(gram_n)}{\sum_{S\in\{Reference\ Summaries\}}\sum_{gram_n\in S} Count(gram_n)} \quad (1)$$

Where *n* is the length of an *n-gram*, gram$_n$ is the maximum number of *n-grams* found in a candidate summary and a set of reference summaries, and count$_{match}$(gram$_n$) is the maximum number of n-grams found in a candidate summary and a set of reference summaries.

ROUGE-L is a statistics program that uses the longest common subsequence, but (LCS). By considering similarities in sentence level structure, the longest common subsequence problem automatically determines the longest co-occurring in sequence n-grams. The benefit of LCS is that it prefers in-sequence matches that reflect sentence level word order rather than consecutive matches. Since it automatically includes the longest common n-grams in sequence, it is not dependent on a specified n-gram length. The ROUGE-N formula can be found in (2).

$$ROUGE - L = \frac{LCS(gram_n)}{Count(gram_n)} \quad (2)$$

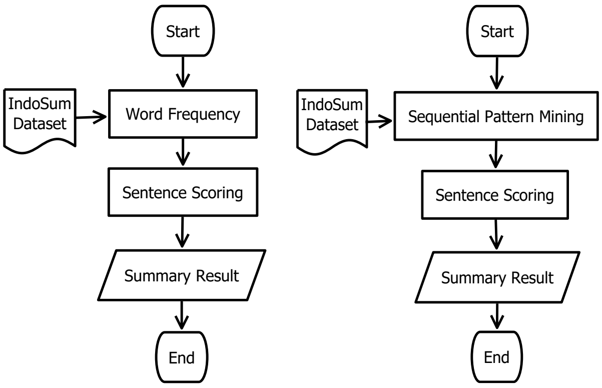
Where LCS(gram$_n$) is longest common subsequence between reference and model output. Then, *gram$_n$* is the maximum number of *n-grams* found in a candidate summary and a set of reference summaries.

3. RESULTS AND DISCUSSION

This section presents the result of the improvement of Indonesian text summary by enhancing the SoW as text representation. It is critical to properly prepare text representation to achieve the readable summary result. This section sequentially presents the process of combination of SPM and sentence scoring, the result of sentence scoring in producing Indonesian text summary, the result of combination of SPM and sentence scoring, and discussion section.

3.1. Combination of sequential pattern mining and sentence scoring process

This section explains the overview of proposed method that combine sentence scoring and SPM in producing Indonesian text summary. Figure 2 illustrate the flow process of feature-based approach using sentence scoring in producing Indonesian text summary. Figure 2(a) shows the sentence scoring process using word frequency, while Figure 2(b) illustrate the combination of SPM and sentence scoring in which SPM replace the word frequency. Sentence Scoring is combined with SPM to produce SoW as text representation. This research uses PrefixSpan algorithm as SPM method that has been used for Indonesian text [22], [38]. The sequence of word that produced from PrefixSpan will replace the word frequency in sentence scoring.



(a) (b)

Figure 2. Illustrate the flow process of feature-based approach using sentence scoring (a) sentence scoring without SPM and (b) sentence scoring with SPM

This research use PrefixSpan algorithm as SPM technique to produce SoW. PrefixSpan or Prefix-projected sequential pattern mining is an SPM algorithm that adapts divide and conquer principles and pattern building to generate efficient sequence patterns from large sequence databases [39]. Here is the PrefixSpan Algorithm 1.

```
Algorithm 1: PrefixSpan
Input: A sequence database S, and the minimum support (threshold min_support)
Output: The complete set of sequential patterns
Method: Call PrefixSpan({},0,S)
Subroutine: PrefixSpan(α,l,$S|_{\alpha}$)
The parameters are 1) α is a sequential pattern; 2) l is the length of α; and 3) $S|_{\alpha}$ is the
α-projected database if α ≠ {}, otherwise, it is the sequence database S.
Method:
1. Scan $S|_{\alpha}$ once, find each frequent item, b, such that
(a) b can be assembled to the last element of α to form a sequential pattern; or
(b) {b} can be appended to α to form a sequential pattern.
2. For each frequent item b, append it to α to form a sequential pattern α′, and output α′.
3. For each α′, construct α′-projected database $S|_{\alpha}$, and call PrefixSpan(α,l+1,$S|_{\alpha'}$)
```

Based on PrefixSpan algorithm that explained above, PrefixSpan will find length-1 sequential pattern first, then dividing the search space. Then, finding subsets of sequential patterns with prefix, respectively. The collection of patterns discovered during the recursive mining process is known as the set of sequential patterns. The illustration of combination between SPM and sentence scoring and Bellman-Ford algorithm will explain in detail in this chapter with the following news text example:

In Indonesian: *Jakarta, CNN Indonesia -- Gubernur Jawa Barat Ridwan Kamil meminta kepolisian untuk memperketat pintu masuk di wilayah perbatasan, termasuk jalan-jalan tikus. Menurut Ridwan Kamil, Pembatasan Sosial Berskala Besar (PSBB) di wilayah Bandung raya yang sudah diberlakukan sejak Rabu (22/4), masih ditemui sejumlah pelanggaran. Salah satu titik yang harus diperbaiki adalah wilayah perbatasan."Mulai sekarang kita perketat penjagaan di perbatasan, tidak boleh ada warga yang masuk maupun keluar dari wilayahnya, kecuali dengan alasan yang jelas," kata pria yang karib disapa Emil itu saat menggelar pertemuan video conference dari Gedung Pakuan, Kota Bandung, Sabtu (25/4).*

In English: Jakarta, CNN Indonesia - West Java Governor Ridwan Kamil has asked the police to tighten entrances in the border area, including rat roads. According to Ridwan Kamil, Large-scale Social Restrictions (PSBB) in the Bandung area, which had been imposed since Wednesday (22/4), several violations were still encountered. One of the points that need to be improved is the border area. "From now on, we tighten security at the border, no residents may enter or leave the area, except for obvious reasons," said the close friend Emil when holding a video conference meeting from Pakuan Building, Bandung City, Saturday (25/4).

Before news text is processed in sequential pattern layer, the news text is prepared and cleaned at the pre-processing stage, including case folding, removing regular expression, removing stop words, stemming, and also separating sentence for the needs of the sentence classification to be selected in the result of summary. From the news text example above, the following results are obtained for text pre-processing and sequential patterns (form minimum of 25% of minimum support or threshold value) as shown in:
<jakarta cnn indonesia>
<gubernur jawa barat ridwan kamil pinta polisi ketat pintu masuk wilayah batas masuk jalan jalan tikus>
<turut ridwan kamil batas sosial skala besar psbb wilayah bandung raya sudah laku sejak rabu masih temu jumlah langgar>
<salah satu titik harus baik wilayah batas>
<mulai ketat jaga batas boleh ada warga masuk mau keluar wilayah alas jelas kata pria karib sapa emil gelar temu video conference gedung pakuan kota bandung sabtu>

The example of Sequential Pattern that produced:
<{ketat}>: 2
<{temu}>: 2
<{masuk}>: 3
<{wilayah}>: 4
<{batas}>: 4
<{ketat, masuk}>: 2
<{ketat, wilayah}>: 2
<{masuk, wilayah}>: 2
<{batas, wilayah}>: 2
<{wilayah, batas}>: 2
<{ketat, masuk, wilayah}>: 2
Etc.

3.2. Combination of sequential pattern mining and sentence scoring process

There are many features that used in sentence scoring method. These characteristics have been used in a variety of strategies, either alone or in conjunction with other techniques. There are at least 22 features used in sentence scoring, including [40] terms like: word similarity among sentences, cue word, title similarity, proper noun, word co-occurance, font style, lexical similarity, term frequency, sentence location, TF-IDF, sentence similarity, numerical value, TextRank, sentence length, positive keyword, negative keyword, busy path, aggregate similarity, word similarity among paragraphs, iterative query. This research use term frequency (TF) or word for sentence scoring method. Term frequency is the first time it was proposed for extracting sentences from text documents. Sentences were graded based on the frequency of terms in the sentence. Term frequency has a greater impact on final scoring.

Term frequency, term in this research is word. So, word frequency is used to extract each word from document and the frequency of occurrence in the document is calculated. This study makes use of normalization TF, which compares a term's frequency to the highest value for the entire or a group of terms in a document. As shown in (3) shows calculation of term frequency.

$$TF = \frac{Word_{doc}}{Max\_Word_{doc}} \quad (3)$$

Where $Word_{doc}$ means the word frequency that appears in the document, while Max_Word_{doc} is maximum frequency word that appears in the document.

Sentence scoring method in this research is a feature-based method (basic method for text summarization) that use word frequency. Where, each word receives a score and the weight of each sentence is the sum of all scores of its constituent words [41], [42]. So, the algorithm begins from:

- Calculate a word frequency to see how often it appears in the text. In other words, the final summary is more likely to include sentences that contain the most common words in the document. The idea is that a word's likelihood of indicating the text's subject increases with its frequency in the text.
- Tokenize the sentences into words by looping through each sentence in the collection of sentences.
- If the sentence is missing from the sentence list, the weighted frequency of the first word in the sentence should be used as the sentence's value. On the other hand, if the sentence already exists, then increase the value by the word's weighted frequency.
- Sentence score in (4).

$$sentence_{score} = \sum word\_frequency\_in\_sentence \quad (4)$$

The example of sentence scoring method with the example document is available:

- Separating sentence (S) and text pre-processing:
  S1: <jakarta, cnn, indonesia>
  S2: <gubernur, jawa, barat, ridwan, kamil, pinta, polisi, ketat, pintu, masuk, wilayah, batas, masuk, jalan, jalan, tikus>
  S3: <turut, ridwan, kamil, batas, sosial, skala, besar, psbb, wilayah, bandung, raya, sudah, laku, sejak, rabu, masih, temu, jumlah, langgar>
  S4: <salah, satu, titik, harus, baik, wilayah, batas>
  S5: <mulai, ketat, jaga, batas, boleh, ada, warga, masuk, mau, keluar, wilayah, alas, jelas, kata, pria, karib, sapa, emil, gelar, temu, video, conference, gedung, pakuan, kota, bandung, sabtu>

- Term/word frequency:

jakarta: 1 -> ¼ = 0.25
cnn: 1 -> ¼ = 0.25
indonesia: 1 -> ¼ = 0.25
gubernur: 1 -> ¼ = 0.25
jawa: 1 -> ¼ = 0.25
barat: 1 -> ¼ = 0.25
ridwan: 2 -> 2/4 = 0.5
kamil: 2 -> 2/4 = 0.5
pinta: 1 -> ¼ = 0.25
polisi: 1 -> ¼ = 0.25
ketat: 1 -> ¼ = 0.25
pintu: 1 -> ¼ = 0.25
masuk: 2 -> 2/4 = 0.5
wilayah: 4 -> 4/4 = 1
batas: 4 -> 4/4 = 1

jalan: 2 -> 2/4 = 0.5
tikus: 1 -> ¼ = 0.25
turut: 1 -> ¼ = 0.25
sosial: 1 -> ¼ = 0.25
skala: 1 -> ¼ = 0.25
besar: 1 -> ¼ = 0.25
psbb: 1 -> ¼ = 0.25
bandung: 2 -> 2/4 = 0.5
raya: 1 -> ¼ = 0.25
sudah: 1 -> ¼ = 0.25
laku: 1 -> ¼ = 0.25
sejak: 1 -> ¼ = 0.25
rabu: 1-> ¼ = 0.25
masih: 1 -> ¼ = 0.25
temu: 2 -> 2/4 = 0.5

jumlah: 1 -> ¼ = 0.25
langgar: 1 -> ¼ = 0.25
salah: 1 -> ¼ = 0.25
satu: 1 -> ¼ = 0.25
titik: 1 -> ¼ = 0.25
harus: 1 -> ¼ = 0.25
baik: 1 -> ¼ = 0.25
mulai: 1 -> ¼ = 0.25
ketat: 2 -> ½ = 0.5
jaga: 1 -> ¼ = 0.25
boleh: 1 -> ¼ = 0.25
ada: 1 -> ¼ = 0.25
warga: 1 -> ¼ = 0.25
mau: 1 -> ¼ = 0.25
keluar: 1 -> ¼ = 0.25

alas: 1 -> ¼ = 0.25
jelas: 1 -> ¼ = 0.25
kata: 1 -> ¼ = 0.25
pria: 1 -> ¼ = 0.25
karib: 1 -> ¼ = 0.25
sapa: 1 -> ¼ = 0.25
emil: 1 -> ¼ = 0.25
gelar: 1 -> ¼ = 0.25
video: 1 -> ¼ = 0.25
conference: 1 -> ¼ = 0.25
gedung: 1 -> ¼ = 0.25
pakuan: 1 -> ¼ = 0.25
kota: 1 -> ¼ = 0.25
sabtu: 1 -> ¼ = 0.25

- Sentence scoring:
  S1: 0.25+0.25+0.25 = 0.75
  S2: 0.25+0.25+0.5+0.5+0.25+0.25+0.25+0.5+0.5+0.5+0.5+0.25=5.0
  S3: 0.25+0.5+0.5+1+0.25+0.25+0.25+0.25+1+0.5+0.25+0.25+0.25+0.25+0.25+0.25+0.5+0.25+0.25= 7.25
  S4: 0.25+0.25+0.25+0.25+0.25+1+1=3.25
  S5:0.25+0.25+0.25+1+0.25+0.25+0.25+0.5+0.25+0.25+1+0.25+0.25+0.25+0.25+0.25+0.25+0.25+ 0.25+0.25+0.25+0.25+0.25+0.25+0.25+0.5+0.25=8.75

- Summary result:

Sentence scoring method will produce the summary based on the maximum number of sentences that want to generate. For example from news text in the subsection 3.1, if maximum sentences in summary is two, then the result of summary is sentence three (S3) and sentence five (S5). The result is "*Menurut Ridwan Kamil, Pembatasan Sosial Berskala Besar (PSBB) di wilayah Bandung raya yang sudah diberlakukan sejak Rabu (22/4), masih ditemui sejumlah pelanggaran. "Mulai sekarang kita perketat penjagaan di perbatasan, tidak boleh ada warga yang masuk maupun keluar dari wilayahnya, kecuali dengan alasan yang jelas," kata pria yang karib disapa Emil itu saat menggelar pertemuan video conference dari Gedung Pakuan, Kota Bandung, Sabtu (25/4).*" In English: (According to Ridwan Kamil, Large-scale Social Restrictions (PSBB) in the Bandung area, which had been imposed since Wednesday (22/4), several violations were still encountered. One of the points that need to be improved is the border area. "From now on, we tighten security at the border, no residents may enter or leave the area, except for obvious reasons," said the close friend Emil when holding a video conference meeting from Pakuan Building, Bandung City, Saturday (25/4)).

3.3. Combination of sequential pattern mining and sentence scoring process

The experiment of proposed method which combine SPM with sentence scoring and Bellman-Ford is conducted using Python with 18,774 news documents that open access and available (named IndoSum) [26]. The news articles dataset are classified by 6 categories: entertainment, inspiration, sport, showbiz, headline, and technology. However, this experiment do not need the categorization because just need to know the performance of proposed method. IndoSum also provide the manual summary to be compared with the summary that produced by system. The experiment contain the text pre-processing including separating sentences, tokenizing, lowering case, removing character non-letter and regular expression, removing Indonesian stopwords, and stemming process using Porter algorithm for Indonesian language. Stop-words removing and stemming process are used Sastrawi library that commonly used for NLP research with

Indonesian text [35]. Then, for PrefixSpan algorithm as a part of SPM method uses prefixspan library for Python [43]. There are two experiments for this section as depicted in Figure 2: i) produce automatic summary with sentence scoring; and ii) produce automatic summary with combination of sentence scoring and SPM. Then, all of the result of experiment is evaluated using ROUGE evaluation metrics. ROUGE evaluation in this experiment uses rouge-score library for Python [44]. Last, the result of experiment will be interpreted and analyzed in the discussion section.

To evaluate the performance of proposed method which combine SPM with sentence scoring, the evaluation process measures the result of summary using ROUGE-1, ROUGE-2 and ROUGE-L. ROUGE evaluation shows the performance of summary result with precision, recall, and f-measure value. Table 1 and Figure 3 show the average value of ROUGE-1, ROUGE-2, and ROUGE-L evaluation of each method.

According to the ROUGE-1 evaluation results, the combination of sentence scoring and SPM has the highest precision, the combination of Bellman-Ford and SPM has the highest recall, and the combination of Bellman-Ford and SPM has the highest f-measure. These findings indicate that when SPM was combined with Sentence Scoring, the precision value increased. The ROUGE-2 result, like the ROUGE-1 result, indicates that SPM can improve the precision of sentence scoring results. The ROUGE-L evaluation yields a very different result. The combination of sentence scoring and SPM improved all metrics, including recall and f-measure.

Sentence scoring with IndoSum dataset yields an average f-measure of 0.44 for ROUGE-1, 0.34 for ROUGE-2, and 0.32 for ROUGE-L. When sentence scoring is combined with SPM, the f-measure of ROUGE-1 and ROUGE-2 drops to 0.40 and 0.33, respectively, while the f-measure of ROUGE-L rises to 0.36. ROUGE-L, which measures long common subsequence (LCS), is more appropriate for the proposed method with SPM. Because SPM generates sequence patterns that pay attention to the order in which words appear, as well as an assessment using ROUGE-L that pays attention to the order of occurrence of words (in the form of n-grams) in a sequence pattern, there is no need to define the n-gram length at the start because it will be calculated automatically in the longest sequence pattern. As a result, while SPM did not improve the f-measures of ROUGE-1 and ROUGE-2, it did improve the f-measure of ROUGE-L, which is a better metric for representing summary quality.

Previous research used the modified SumBasic method with the same IndoSum dataset in their experiment [26]. SumBasic is a feature-based approach that ranks sentences based on word frequency and chooses the top sentences to serve as the summary [12], [45]. Figure 4 and Table 2 provide the comparison of ROUGE score between SumBasic, sentence scoring, and combination of sentence scoring and SPM. SumBasic's f-measure from ROUGE-1 is 0.359, ROUGE-2 is 0.202, and ROUGE-L is 0.338. When compared to the performance of the proposed method, the f-measure results of ROUGE-1, ROUGE-2, and ROUGE-L are higher than SumBasic. This means that sentence scoring combined with SPM outperforms SumBasic. Furthermore, when sentence scoring is combined with SPM, the f-measure of ROUGE-L increases.

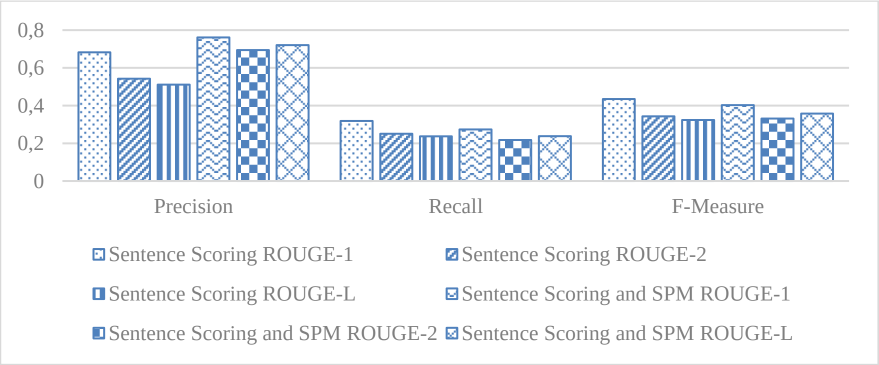


Figure 3. ROUGE evaluation result of sentence scoring and proposed method

Table 1. ROUGE evaluation result of sentence scoring and proposed method

Method	Evaluation	Precision	Recall	F-Measure
Sentence scoring	ROUGE-1	0.6825	0.3187	0.4345
	ROUGE-2	0.5420	0.2506	0.3427
	ROUGE-L	0.5112	0.2370	0.3238
Sentence scoring and SPM	ROUGE-1	**0.7615**	0.2737	0.4026
	ROUGE-2	**0.6946**	0.2177	0.3315
	ROUGE-L	**0.7205**	0.2377	**0.3575**

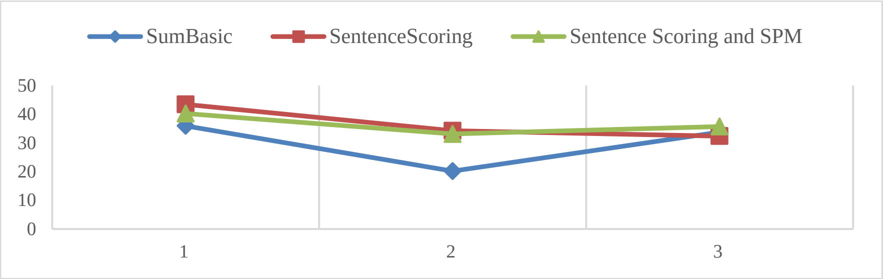


Figure 4. Comparison of ROUGE evaluation result between proposed method and SumBasic

Table 2. Comparison of ROUGE evaluation result between proposed method and SumBasic

Method	ROUGE Evaluation (Percentage of F-Measure)		
	ROUGE-1	ROUGE-2	ROUGE-L
SumBasic [26]	35.96	20.19	33.77
Sentence scoring	43.45	34.27	32.38
Sentence scoring and SPM	40.26	33.15	**35.75**

4. CONCLUSION

This research aims to enhance the quality of Indonesian automatic text summary in the aspect of readability. This chapter provides the study on sequential pattern mining (SPM) that combined with feature-based approach, such as sentence scoring. SPM replace the word-frequency calculation in sentence scoring. The experiment is conducted using IndoSum dataset with four scenarios: producing Indonesian text summary using sentence scoring and producing Indonesian text summary using combination of sentence scoring and SPM. Then, the performance of proposed method is evaluated with co-selection-based analysis using ROUGE-1, ROUGE-2, and ROUGE-L. Based on the ROUGE evaluation result, overall SPM can improve the quality of summary result compared without using SPM, especially in precision and f-measure value. Precision value of ROUGE-1, ROUGE-2, and ROUGE-L of combination between sentence scoring and SPM is better than sentence scoring only. Although recall and f-measure at ROUGE-1 and ROUGE-2 in the combination of sentence scoring and SPM did not increase, but, at ROUGE-L both precision, recall, and f-measure increased. Seeing these results, combining SPM in the automatic text summary process has the potential to be further enhanced to improve Indonesian text summary performance. For the further works, SPM can be combined with the other methods to enhance the quality of Indonesian automatic text summarization. Then, it is important to evaluate the readability of summary result, not only using co-selection-based analysis, but also using content-based analysis and involve an Indonesian language expert to evaluate the summary result.

ACKNOWLEDGEMENTS

The author would like to thank for the support from the academic community of UIN Sunan Gunung Djati Bandung, especially Prof. Dr. H. Muhammad Ali Ramdhani, S.TP., MT for his advice and knowledge.

REFERENCES

[1] A. Kumar, Z. Luo, and M. Xu, “Text summarization using natural language processing,” *Worcester Polytechnic Institute*, 2018. [Online]. Available: https://web.wpi.edu/Pubs/E-project/Available/E-project-042418-172843/unrestricted/juniper_final_report.pdf.

[2] G. G. Chowdhury, “Natural language processing,” *Annual Review of Information Science and Technology*, vol. 37, no. 1, pp. 51–89, Jan. 2005, doi: 10.1002/aris.1440370103.

[3] J. Hirschberg and C. D. Manning, “Advances in natural language processing,” *Science*, vol. 349, no. 6245, pp. 261–266, Jul. 2015, doi: 10.1126/science.aaa8685.

[4] P. M. Nadkarni, L. Ohno-Machado, and W. W. Chapman, “Natural language processing: an introduction,” *Journal of the American Medical Informatics Association*, vol. 18, no. 5, pp. 544–551, Sep. 2011, doi: 10.1136/amiajnl-2011-000464.

[5] M. Gambhir and V. Gupta, “Recent automatic text summarization techniques: a survey,” *Artificial Intelligence Review*, vol. 47, no. 1, pp. 1–66, Jan. 2017, doi: 10.1007/s10462-016-9475-9.

[6] V. Gupta and G. S. Lehal, “A Survey of Text Summarization extractive techniques,” *Journal of Emerging Technologies in Web Intelligence*, vol. 2, no. 3, pp. 258–268, Aug. 2010, doi: 10.4304/jetwi.2.3.258-268.

[7] M. Rajangam and C. Annamalai, “Extractive document summarization using an adaptive, knowledge based cognitive model,” *Cognitive Systems Research*, vol. 56, pp. 56–71, Aug. 2019, doi: 10.1016/j.cogsys.2018.11.005.

[8] N. R. Kasture, N. Yargal, N. N. Singh, N. Kulkarni, and V. Mathur, “A survey on methods of abstractive text summarization,” *International Journal for Research in Emerging Science and Technology*, vol. 7, no. 1, pp. 728–734, 2014.

[9] J. Jayabharathy, S. Kanmani, and Buvana, "Multi-document summarization based on sentence features and frequent itemsets," in *Advances in Intelligent and Soft Computing*, vol. 166 AISC, no. VOL. 1, 2012, pp. 657–671.

[10] T. Sri, R. Raju, and B. Allarpu, "Text summarization using sentence scoring method," *International Research Journal of Engineering and Technology*, vol. 4, no. 4, pp. 1777–1779, 2017, [Online]. Available: https://irjet.net/archives/V4/i5/IRJET-V4I5493.pdf.

[11] P. M. Sabuna and D. B. Setyohadi, "Summarizing Indonesian text automatically by using sentence scoring and decision tree," in *Proceedings - 2017 2nd International Conferences on Information Technology, Information Systems and Electrical Engineering, ICITISEE 2017*, Nov. 2018, vol. 2018-January, pp. 1–6, doi: 10.1109/ICITISEE.2017.8285473.

[12] L. Vanderwende, H. Suzuki, C. Brockett, and A. Nenkova, "Beyond SumBasic: Task-focused summarization with sentence simplification and lexical expansion," *Information Processing & Management*, vol. 43, no. 6, pp. 1606–1618, Nov. 2007, doi: 10.1016/j.ipm.2007.01.023.

[13] M. G. Ozsoy, F. N. Alpaslan, and I. Cicekli, "Text summarization using latent semantic analysis," *Journal of Information Science*, vol. 37, no. 4, pp. 405–417, Aug. 2011, doi: 10.1177/0165551511408848.

[14] K. Merchant and Y. Pande, "NLP based latent semantic analysis for legal text summarization," in *2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, Sep. 2018, pp. 1803–1807, doi: 10.1109/ICACCI.2018.8554831.

[15] J. Steinberger and K. Ježek, "Using latent semantic analysis in text summarization," in *In Proceedings of ISIM 2004*, 2004, pp. 93--100.

[16] Y. J. Kumar, O. S. Goh, H. Basiron, N. H. Choon, and P. C. Suppiah, "A review on automatic text summarization approaches," *Journal of Computer Science*, vol. 12, no. 4, pp. 178–190, 2016, doi: 10.3844/jcssp.2016.178.190.

[17] J. ge Yao, X. Wan, and J. Xiao, "Recent advances in document summarization," *Knowledge and Information Systems*, vol. 53, no. 2, pp. 297–336, Nov. 2017, doi: 10.1007/s10115-017-1042-4.

[18] K. Nandhini and S. R. Balasundaram, "Improving readability through extractive summarization for learners with reading difficulties," *Egyptian Informatics Journal*, vol. 14, no. 3, pp. 195–204, 2013, doi: 10.1016/j.eij.2013.09.001.

[19] P. Verma and H. Om, "MCRMR: Maximum coverage and relevancy with minimal redundancy based multi-document summarization," *Expert Systems with Applications*, vol. 120, pp. 43–56, Apr. 2019, doi: 10.1016/j.eswa.2018.11.022.

[20] P. Verma, S. Pal, and H. Om, "A comparative analysis on hindi and english extractive text summarization," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 18, no. 3, pp. 1–39, Sep. 2019, doi: 10.1145/3308754.

[21] P. Verma and H. Om, "A novel approach for text summarization using optimal combination of sentence scoring methods," *Sādhanā*, vol. 44, no. 5, p. 110, May 2019, doi: 10.1007/s12046-019-1082-4.

[22] D. Rahmawati, G. A. Putri Saptawati, and Y. Widyani, "Document clustering using sequential pattern (SP): Maximal frequent sequences (MFS) as SP representation," in *Proceedings of 2015 International Conference on Data and Software Engineering, ICODSE 2015*, Nov. 2016, pp. 98–102, doi: 10.1109/ICODSE.2015.7436979.

[23] S. Alias, S. K. Mohammad, K. H. Gan, and T. T. Ping, "MYTextSum: A Malay text summarizer model using a constrained pattern-growth sentence compression technique," in *Lecture Notes in Electrical Engineering*, vol. 488, 2018, pp. 141–150.

[24] S. Alias, S. K. Mohammad, G. K. Hoon, and T. T. Ping, "A Malay text summarizer using pattern-growth method with sentence compression rules," in *2016 Third International Conference on Information Retrieval and Knowledge Management (CAMP)*, Aug. 2016, pp. 7–12, doi: 10.1109/INFRKM.2016.7806326.

[25] J. Pei *et al.*, "Mining sequential patterns by pattern-growth: The prefixspan approach," *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, no. 11, pp. 1424–1440, Nov. 2004, doi: 10.1109/TKDE.2004.77.

[26] K. Kurniawan and S. Louvan, "Indosum: A new benchmark dataset for Indonesian text summarization," in *2018 International Conference on Asian Language Processing (IALP)*, Nov. 2018, pp. 215–220, doi: 10.1109/IALP.2018.8629109.

[27] F. Koto, J. H. Lau, and T. Baldwin, "Liputan6: A Large-scale Indonesian dataset for text summarization," *arXiv preprints*, Nov. 2020, doi: 10.48550/arXiv.2011.00679.

[28] S. Cahyawijaya *et al.*, "IndoNLG: Benchmark and Resources for Evaluating Indonesian Natural Language Generation," in *EMNLP 2021 - 2021 Conference on Empirical Methods in Natural Language Processing, Proceedings*, Apr. 2021, pp. 8875–8898, doi: 10.18653/v1/2021.emnlp-main.699.

[29] N. Lin, J. Li, and S. Jiang, "A simple but effective method for Indonesian automatic text summarisation," *Connection Science*, vol. 34, no. 1, pp. 29–43, Dec. 2022, doi: 10.1080/09540091.2021.1937942.

[30] D. Fitrianah and R. N. Jauhari, "Extractive text summarization for scientific journal articles using long short-term memory and gated recurrent units," *Bulletin of Electrical Engineering and Informatics*, vol. 11, no. 1, pp. 150–157, Feb. 2022, doi: 10.11591/eei.v11i1.3278.

[31] R. Wijayanti, M. L. Khodra, and D. H. Widyantoro, "Indonesian Abstractive Summarization using Pre-Trained Model," in *3rd 2021 East Indonesia Conference on Computer and Information Technology, EIConCIT 2021*, Apr. 2021, pp. 79–84, doi: 10.1109/EIConCIT50028.2021.9431880.

[32] F. Halim, L. Liliana, and K. Gunadi, "Automatic extractive summaries on Indonesian language news using the BERT method (in Indonesia: *Ringkasan Ekstraktif Otomatis pada Berita Berbahasa Indonesia Menggunakan Metode BERT*)," *Jurnal Infra*, vol. 10, no. 1, pp. 162–168, 2022.

[33] K. Kurniawan and S. Louvan, "Indosum Dataset," Gitub, 2019. https://github.com/kata-ai/indosum (accessed Nov. 06, 2020).

[34] S. Vijayarani, M. J. Ilamathi, M. Nithya, A. Professor, and M. P. Research Scholar, "Preprocessing techniques for text mining -an overview," *International Journal of Computer Science & Communication Networks*, vol. 5, no. 1, pp. 7–16, 2015.

[35] H. A. Robbani, "Sastrawi Python," *Github Repository*, 2018. https://github.com/har07/PySastrawi (accessed Jan. 30, 2019).

[36] J. Asian, H. E. Williams, and S. M. M. Tahaghoghi, "Stemming Indonesian," *Conferences in Research and Practice in Information Technology Series*, vol. 38, no. 4, pp. 307–314, Dec. 2005, doi: 10.1145/1316457.1316459.

[37] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Proceedings of the workshop on text summarization branches out (WAS 2004)*, 2004, pp. 74–81.

[38] D. Sa'adillah Maylawati, M. Irfan, and W. Budiawan Zulfikar, "Comparison between BIDE, PrefixSpan, and TRuleGrowth for Mining of Indonesian Text," *Journal of Physics: Conference Series*, vol. 801, no. 1, p. 012067, Jan. 2017, doi: 10.1088/1742-6596/801/1/012067.

[39] J. Han, J. Pei, Y. Yin, and R. Mao, "Mining frequent patterns without candidate generation: A frequent-pattern tree approach," *Data Mining and Knowledge Discovery*, vol. 8, no. 1, pp. 53–87, Jan. 2004, doi: 10.1023/B:DAMI.0000005258.31418.83.

[40] Y. K. Meena and D. Gopalani, "Analysis of sentence scoring methods for extractive automatic text summarization," in *Proceedings of the 2014 International Conference on Information and Communication Technology for Competitive Strategies*, Oct. 2014, vol. 11-16-Nove, pp. 1–6, doi: 10.1145/2677855.2677908.

[41] R. Ferreira *et al.*, "Assessing sentence scoring techniques for extractive text summarization," *Expert Systems with Applications*, vol. 40, no. 14, pp. 5755–5764, 2013.
[42] U. Malik, "Text summarization with NLTK in Python," *Retrieved from Stack Abuse*, 2018 (accessed Feb. 15, 2019).
[43] C. Gao, "prefixspan 0.5.2," *pypi.org*, 2018. https://pypi.org/project/prefixspan/ (accessed Jul. 01, 2019).
[44] Google LLC, "rouge-score 0.1.2," *pypi.org*, 2021. https://pypi.org/project/rouge-score/ (accessed Sep. 20, 2021).
[45] A. Nenkova and L. Vanderwende, "The impact of frequency on summarization: MSR-TR-2005-101," *Redmond, Washington: Microsoft Research*, vol. 101, pp. 1–9, 2005, [Online]. Available: http://duc.nist.gov.

BIOGRAPHIES OF AUTHORS

Dian Sa'adillah Maylawati is a lecturer in Department of Informatics at UIN Sunan Gunung Djati Bandung, Indonesia. She got a Ph.D. from Centre for Advanced Computing Technology, Faculty of Information and Communication Technology in Universiti Teknikal Malaysia Melaka (UTeM), Malaysia. She got a master's degree from Bandung Technology Institute, Indonesia. Her current research interests focus on software engineering, machine learning, text mining, and natural language processing. She can be contacted at email: diansm@uinsgd.ac.id.

Dr. Yogan Jaya Kumar is a lecturer at Centre for Advanced Computing Technology, Faculty of Information and Communication Technology in Universiti Teknikal Malaysia Melaka (UTeM), Malaysia. He got a doctoral degree from Universiti Teknologi Malaysia Johor. His current position is a Centre for Advanced Computing Technology, Faculty of Information and Communication Technology in Universiti Teknikal Malaysia Melaka (UTeM), Malaysia. His current research interests focus on soft computing, text mining, and information retrieval. He can be contacted at email: yogan@utem.edu.my.

Dr. Fauziah Binti Kasmin is a senior lecturer at Centre for Advanced Computing Technology, Faculty of Information and Communication Technology in Universiti Teknikal Malaysia Melaka (UTeM), Malaysia. She got doctoral degree from Universiti Kebangsaan Malaysia. Her current research interest focuses on statistics and image processing. She can be contacted at email: fauziah@utem.edu.my.