

## Feature-based POS tagging and sentence relevance for news multi-document summarization in Bahasa Indonesia

Moch Zawaruddin Abdullah<sup>1</sup>, Chastine Fatichah<sup>2</sup>

<sup>1</sup>Department of Information Technology, Politeknik Negeri Malang, Malang, Indonesia

<sup>2</sup>Department of Informatics, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia

---

### Article Info

#### Article history:

Received Aug 25, 2020

Revised Mar 30, 2021

Accepted Dec 4, 2021

---

#### Keywords:

Indonesian news

Multi-document summarization

News features

Pos tagging

Sentence extraction

Sentence relevance

---

### ABSTRACT

Sentence extraction in news document summarization determines representative sentences primarily by employing the news feature known as news feature score (NeFS). NeFS can achieve meaningful sentences by analyzing the frequency and similarity of phrases while neglecting grammatical information and sentence relevance to the title. The presence of instructive content is indicated by grammatical information carried by part of speech (POS). POS tagging is the process of giving a meaningful tag to each term based on qualified data and even surrounding words. Sentence relevance to the title is intended to determine the sentence's level of connectivity to the title in terms of both word-based and meaning-based similarity, primarily for news documents in Bahasa Indonesia. In this study, we present an alternative sentence weighting method by incorporating news features, POS tagging, and sentence relevance to the title. Sentence extraction based on news features, POS tagging, and sentence relevance is introduced to extract the representative sentences. The experiment results on the 11 groups of Indonesian news documents are compared with the news features scores with the grammatical information approach method (NeFGIS). The proposed method achieved better results. The increasing f-score rate of ROUGE-1, ROUGE-2, ROUGE-L, and ROUGE-SU4 sequentially are 1.84%, 3.03%, 3.85%, 2.08%.

*This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.*



---

### Corresponding Author:

Moch Zawaruddin Abdullah

Department of Information Technology, State Polytechnic of Malang

9 Soekarno Hatta St., Malang, East Java, Indonesia

Email: [zawaruddin@polinema.ac.id](mailto:zawaruddin@polinema.ac.id)

---

## 1. INTRODUCTION

A summary can be understood as a text containing a few phrases that convey a document's essential information. It presents the crucial concept of an article's substance, aimed to tell readers the main idea of the text [1]. The length of a summary is no more than half the original document's length and is usually shorter [2]. Multi-document summarization generates a representative summary of the entire document by reducing size forms while retaining the original essential information [3]. It takes considerable time to locate a representative sentence from all documents. For example, sentences known in this document as primary sentences are classified as descriptive sentences and as candidates for summary phrases.

Sentence weighting is intended to select crucial sentences as the basis of the summary. The important sentences should contain as much detail as possible [4] and have essential terms from the source text [5]. In the news document summaries, the weighting system primarily uses the news function called the news feature scoring (NeFS) [6]. Extractive document summaries are achieved by extracting and sorting sentences through their highest value and using them as sentence candidates of resumes. The extractive

summary words are similar to those contained in the original text [7]. A good summary should provide extensive coverage of the source text and a high coherence interconnection degree [8]. The summary must also include many of the main definitions (saliency) present in the source text [9]. Coverage, coherence, and saliency shall be the benchmarks for choosing representative sentences in news multi-document summaries.

There are diverse methods of sentence weighting for summarizing multi-document news. One involves the use of news features to weigh the sentence, i.e., sentence position, centroid, and sentence resemblance to the news's first sentence [10]. This approach is capable of producing reasonably valid findings in the document summary. However, the method should not pay attention to the relationship between the title and the text such that the sentences chosen for the overview document are less coherent. Another technique uses a variety of news features for sentence weighting. It concludes a compelling mixture of weighting features, i.e., word frequency, TF-IDF, lexical similarity, and sentence length [11]. Another method makes use of a statistical decision by calculating the weight of sentences based on the part of speech tagging (POS tagging) [6], [7], [12], [13]. According to the POS tagging approach, the essential sentences that turn out to be candidates for summary sentences are obtained based on the distribution of important sentences with many frequencies, spread, and most content-bearing words. The news features approach focuses on selecting representative sentences by utilizing news features. It obtains intelligible phrases by detecting frequency and assessing words similarity while ignoring grammatical information. Grammatical information carried by part of speech tagging indicates the presence of informative content. On the other side, the POS tagging approach focuses on finding important sentences by determining sentences with many frequencies, well spread, and most content-bearing words.

Summarization based on news features or POS tagging approach must be improved, specifically for news document summaries in Bahasa Indonesia. The news features approach can produce a good result by using news features. Still, it could be inferior if the frequencies and word spread are fewer than content-bearing words, and it cannot distinguish the different functions or meanings of a word. This problem can be solved by using grammatical information that is carried by a POS label. Furthermore, sentence weighting based on POS tagging ignores sentence relevance to the title of the document. Sentence relevance to the title is intended to determine the sentence's level of connectivity to the title in terms of both word-based and meaning-based similarity. Two sentences are considered similar or relevant if most of the words are the same or if they are a paraphrase of each other [14]. News features with grammatical information and sentence relevance to the title approach can be a great combination to find important sentences. It will have arranged the summary from sentences with the most informative and relevant content to the title.

## 2. RESEARCH METHOD

In this study, the research technique from [6], [11] was adopted, and the framework by [6], [13], was also employed. The three most essential stages are conducted to obtain final summaries: i) text preprocessing, ii) sentence extraction for extracting the representative sentence, and iii) arrangement of summaries from representative sentences.

### 2.1. Text preprocessing phase

Text preprocessing is used for sentence segmentation and term construction to be processed and ensured that the textual content is more structured and compatible with the system. In this study, we use seven steps in text preprocessing:

- Sentences segmentation is the process of decoding the text of a document into a collection of sentences. It is very critical concerning the operation of sentence weighting. Misplacement inside the segmentation can lead to a miscalculation in determining the sentence representative, which makes the summary results inappropriate.
- POS tagging is used to label each word with its POS label of a sentence.
- Case folding is the process of converting all characters of sentences into the same format (lowercase). This case folding improves the accuracy of the system to distinguish similar words.
- Tokenizing is conducted to split the sentence into words so that each word can stand alone.
- Stopword removal is done to eliminate phrases that have less significance in a sentence.
- Stemming is the process to obtain the primary word of each word.
- Sentence length threshold is a threshold that a sentence can be scored. Summaries must avoid a phrase too long or too short.

### 2.2. Sentence extraction phase

Sentence extraction is a method of sentence weighting to determine whether or not sentences are meaningful as summary sentences. This study incorporates and enhances Abdullah's research technique

(NeFGIS) [6] employs a sentence relevance strategy. Figure 1 indicates three main components of sentence extraction: i) news feature with three sub-components used for weighting sentences; local sentence weighting using TF-ISF approach, global sentence weighting using TF-IDF approach, and sentence position, ii) part of speech tagging with sub-components; local distribution and global distribution, iii) sentence relevance to the title with sub-components; n-gram word similarity and query expansion similarity. Sentence relevance to the title is the improvement strategy for NeFGIS in this research.

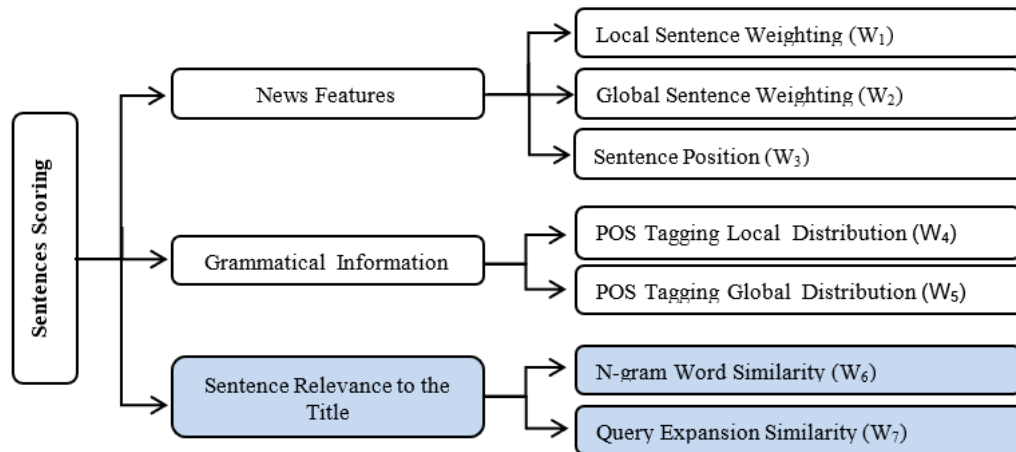


Figure 1. The steps involved in sentence extraction

**2.2.1. News features**

News features score is concerned with selecting representative sentences, which are primarily based on features of the news itself. Based on the research method from [6], there are four scoring methods adopted; local sentence weighting, global sentence weighting, sentence position, resemblance to the title. Furthermore, the sentence resemblance to the title will be used as part of the sentence relevance approach that is the contribution of this research.

a. Local sentence weighting ( $W_1$ )

The local sentence weighting method is a weighting process applied to sentences in a single document. This weighting technique is used to extract sentences instead of documents. It should employ a corpus weighting mechanism based on the amount of sentences rather than the number of documents [15]. This method uses the term frequency-inverse sentence frequency (TF-ISF) approach, such as research conducted by [12], [16] to remove the impact of higher frequency terms, which are not useful. The highest-score sentence is a representative sentence that represents the content of the document. Each term  $t$  will be assigned a term weight ( $w_t$ ) using (1).

$$w_t = TF_t * ISF_t \tag{1}$$

$TF_t$  is the frequency of a particular word in which  $t$  appears in a document, and  $ISF_t$  is represented by (2).

$$ISF_t = \log \left( 1 + \frac{N}{sf(t)} \right) \tag{2}$$

$N$  is the number of sentences in a text document, and  $sf(t)$  is the number of sentences in which the term  $t$  occurred. The local sentence weight of the sentence  $W_1(S_i)$  can be defined as (3).

$$W_1(S_i) = \log \left( 1 + \frac{\sum_{j=1}^k w_t}{k} \right) \tag{3}$$

Where  $k$  is the number of words in a sentence  $i$  ( $S_i$ ).

b. Global sentence weighting ( $W_2$ )

The global sentence weighting method is a weighting process applied to sentences based on words that appear on some documents. The same words that spread over several documents indicate that the words are essential and represent the degree of similarity of documents. Important words in the sentence may

represent some documents in determining the representative sentence for document summarization. This method uses the *term frequency-inverse document frequency* (TF-IDF) approach, such as research conducted by [11] to remove the impact of higher frequency terms, which are not useful in the final summary. Each term  $t$  will be assigned a term weight ( $w_t$ ) using (4).

$$w_t = TF_t * IDF_t \quad (4)$$

$TF_t$  is the frequency of a particular word in which  $t$  appears in the documents, and IDF reflects the terms' distribution within the corpus.  $IDF_t$  is represented by (5).

$$IDF_t = \log \left( 1 + \frac{N}{df(t)} \right) \quad (5)$$

$N$  is the total number of corpus documents, and  $df(t)$  is the document that contains at least one occurrence of the term  $t$ . The global sentence weight of the sentence  $W_2(S_i)$  can be defined as (6).

$$W_2(S_i) = \log \left( 1 + \frac{\sum_{j=1}^k w_t}{k} \right) \quad (6)$$

Where  $k$  is the number of words in a sentence  $i$  ( $S_i$ ). Ideally, the system should assign the highest weight to terms with the most discriminative power [15].

#### c. Sentence position ( $W_3$ )

Adopted from the research [11], [16], [17], the sentence at the beginning of the document has a higher score than that at the last position. It is based on the assertion that most news tends to present the main idea at the beginning of the sentence, while the next sentence is an explanation or even other information outside the subject. The sentence weighting based on sentence position can be explained in (7).

$$W_3(S_i) = \frac{1}{\sqrt{POS(S_i)}} \quad (7)$$

$POS(S_i)$  is the index position of sentence  $S_i$  that appears in the document

### 2.2.2. Grammatical information

Based on Abdullah's research [6], each sentence contains grammatical information that may indicate whether or not the sentence is important to a document. Grammatical information carried by part of speech (POS) indicates to an extent the presence or absence of informative content in a sentence and increases the quality of translation [13], [17]. Part of Speech tagging is assigning a relevant tag to each term based on qualified data and even on surrounding words [18]. POS can provide some information about a word (noun, verb) and the words around it (possessive pronoun, personal pronoun) in natural language processing [19]. POS tag can mark a term that has appeared more often in the document and can be the most important term [20]. Based on Jespersen's Rank Theory, POS can be ranked into four degrees: i) nouns, because they have the most content-bearing labels, ii) adjectives, verbs, and participles, ii) adverbs, and finally iv) all remaining POS [13]. In this study, we make a list of POS label weight  $PW$  based on Jespersen's Rank Theory [21] with four weight values, and our experiment determines the  $PW$  values which can be defined as (8).

$$PW = \{1, 0.75, 0.5, 0.25\} \quad (8)$$

#### a. POS tagging local distribution ( $W_4$ )

The weight of  $W_4$  is calculated for each term  $t$  in sentence  $S_i$ . It calculates the degree of term distribution that has labeled POS  $p$  on a single document. This process is done for each document in the corpus separately.  $W_4$  is calculated for each term  $t$  in the sentence  $i$ . The term  $t$  is a combination of the word with its POS label in the sentence  $i$ . A single term (basic word) can obtain a multi POS label in the document.  $W_4$  can be obtained from (9).

$$W_4(S_i) = \log \left( 1 + \frac{\sum_{j=1}^k (TF_{d,t,p} * PW_p)}{k} \right) \quad (9)$$

$TF_{d,t,p}$  is the frequency of term  $t$  that has labeled POS  $p$  in document  $d$ .  $PW_p$  is POS label weight labeled as  $p$ .  $k$  is the number of the word in a sentence  $i$  ( $S_i$ ).

b. POS tagging global distribution ( $W_5$ )

POS tagging global distribution is used to determine the distribution of terms labeled with the same POS label in all documents. Similar to POS tagging local distribution, the weight of POS tagging global distribution is calculated to each term  $t$  in the sentence  $i$ . The term  $t$  is a combination of a word with its POS label in the sentence  $i$ .  $W_5$  can be obtained from (10).

$$W_5(S_i) = \log \left( 1 + \frac{\sum_{j=1}^k (TF_{t,p} * PW_p)}{k} \right), \quad (10)$$

Where  $TF_{t,p}$  is the frequency of term  $t$  that has labeled POS  $p$  on all documents.  $PW_p$  is POS label weight marked as  $p$ .  $k$  is the number of words in a sentence  $i$  ( $S_i$ ).

### 2.2.3. Sentence relevance to the title

Abdullah [6] and Ferreira [11] used sentence resemblance to the title strategy in order to determine important words. This approach assumes that sentences with a high degree of similarity to the title are essential sentences. It is based on the number of words in the sentence contained in the title. The more terms in the sentence contained in the title, the higher the score attained.

Sentence relevance to the title is our proposed strategy for enhancing the summary result based on sentence resemblance to the title. This method uses sentence weighting to determine the degree of connectivity of a sentence to the title based on two factors; the word-based similarity with n-gram word similarity approach and meaning-based (synonym) similarity with query expansion similarity approach. The use of a mix of word-based and meaning-based similarities on sentence relevancy for selecting essential sentences could improve the quality of the summary that is most relevant to the title.

a. N-gram word similarity ( $W_6$ )

An n-gram is a contiguous sequence of  $n$  words from a given text [22], in which size  $n=1$  (or 1-gram) is referred to as unigram; size 2 is a bigram; size 3 is a trigram; and so on. N-gram word similarity is intended to compare two sentences based on the  $n$ -word sequence. This weighting is based on the resemblance of  $n$  the word sequence between the sentence and the document title. Based on research from [6], [11], sentence resemblance to the title can be obtained by (11).

$$w_t = \frac{NTW}{T} \quad (11)$$

$NTW$  is the number of title words in a sentence, and  $T$  is the number of words in the title. We argue that this weighting performs unigram comparisons by directly finding out the existence of a word from the title in the sentence rather than determining the multi-word expression and the order of words in the text. For example, there are two sentences to be compared; "i love the cat" and "the cat i love". It will result in precisely the same if using unigram. We decided to improve the weighting based on n-gram with  $n$  value as the minimum number of words between sentence  $I$  and the title. N-gram word similarity ( $W_6$ ) can be defined as (12).

$$W_6(S_i) = \log \left( 1 + \frac{\sum_{j=1}^{n-1} \left( \frac{NTW_j}{T_j} \right)}{n} \right) \quad (12)$$

$$n = \min(\text{length}(S_i), \text{length}(\text{title}))$$

Where  $NTW_j$  is the number of word sequences in a sentence  $i$  similar to the word sequence in the title at the  $j$ -gram level.  $T_j$  is the number of word sequences in the title at  $j$ -gram level.  $n$  is the minimum value of the number of words between sentence  $i$  and the title.

b. Query expansion similarity ( $W_7$ )

Query expansion is the process of supplementing additional words or phrases to the original query to improve the retrieval performance [23] using a dictionary or general thesaurus [24]. Query expansion similarity is the meaning-based word similarity between two sentences. In this weighting, the title is considered a query, and expansions are made using the thesaurus. The word's expansion in the title is done to get an alternative word with the same meaning possessed by the sentence in the document. Query expansion similarity ( $W_7$ ) can be obtained by (13).

$$W_7(S_i) = \log \left( 1 + \frac{\sum_{j=1}^n \left( \frac{NTW_{QE,j}}{T_{QE,j}} \right)}{n} \right) \quad (13)$$

Where  $NTW_{QE,j}$  is the number of terms in sentence  $i$ , similar to the query expansion of the term  $j$  in the title.  $T_{QE,j}$  is the number of terms expansion of the term  $j$  in the title.  $n$  is the number of words in the title.  $k$  is the number of words in a sentence  $i$  ( $S_i$ ).

### 3. RESULTS AND ANALYSIS

The experiments in this study were conducted using two-sentence weighting methods for comparison. The news feature and grammatical information approach method (NeFGIS) [6] and our proposed method. The data used in this study is the dataset from [6], comprising 11 topics of news documents in Bahasa Indonesia with three (3) reference summaries for each group. The datasets are presented in Table 1. The evaluation of summary results uses ROUGE-1, ROUGE-2, ROUGE-L, and ROUGE-SU4, where the higher the value, the better the quality is, as shown in [25].

Table 1. The Indonesian news dataset

No	Indonesian topics	English meaning	Total documents	Total sentences
1	Air-Asia	Air-Asia plane	7	88
2	Banjarnegara	Banjarnegara Regency	15	195
3	BBM	fuel	12	204
4	BPJS	Healthcare and social security agency	17	295
5	Dolly	Dolly (place)	9	180
6	Ebola	Ebola	7	86
7	Kurikulum 2013	Education curriculum 2013	23	403
8	Palestina	Palestine	17	186
9	Pilpres	Presidential election	18	231
10	Sinabung	Mount Sinabung	6	87
11	U19	Under-19 football national team	9	142
Total:			140	2097

In our proposed method, experiments were performed using two-sentence selecting techniques, top global sentences and top local sentences. The top global sentence is sentence retrieval that has been sorted by the highest score on one topic. The top local sentence is sentence retrieval by selecting one sentence with the highest score on one topic in each document. The result of the sentence selection for the proposed method is shown in Table 2.

Table 2. The experiment of the proposed method using two-sentence selecting techniques

Rouge	Experiment	Recall	Precision	F-Score
Rouge-1	Global	0.29972	0.62563	0.40361
	Local	0.29888	0.61405	0.39989
Rouge-2	Global	0.15528	0.36406	0.21646
	Local	0.15010	0.34880	0.20835
Rouge-L	Global	0.28956	0.52154	0.37103
	Local	0.29469	0.52266	0.37597
Rouge-SU4	Global	0.17714	0.40957	0.24597
	Local	0.17411	0.39866	0.24060

Table 2 shows that the top global sentence is superior in Rouge-1, Rouge-2, Rouge-SU4 evaluations, and the top local sentence is outstanding in Rouge-L evaluation. The top global sentence is superior to the top local sentences because the news on one topic has a high resemblance. In summary, the sentence between documents with a high similarity is considered one sentence representing the documents. Consequently, the top global sentence technique will be used in the testing to determine the result of the proposed method compared with NeFGIS [6]. The result of testing for the proposed method and NeFGIS about "pilpres" (Presidential election) is shown in Table 3.

Table 3. Summary evaluation result for “*pilpres*” (Presidential election) topic

Summary method	Rouge	Recall	Precision	F-Score
Proposed method	Rouge-1	0.31599	0.53086	0.39617
	Rouge-2	0.17605	0.31099	0.22482
	Rouge-L	0.32992	0.50694	0.39971
	Rouge-SU4	0.19183	0.34398	0.24631
NeFGIS method	Rouge-1	0.32790	0.47810	0.38900
	Rouge-2	0.17700	0.28440	0.21820
	Rouge-L	0.34180	0.44060	0.38490
	Rouge-SU4	0.19590	0.31380	0.24130

Table 3 reveals that the NeFGIS outperforms the suggested technique in terms of recall for all Rouge assessments, but the proposed method outperforms it in terms of precision and f-score. NeFGIS is good at recall because, in the sentence resemblance to the title approach, it counts the number of words in the sentence contained in the title. The higher the score, the more terms in the sentence contained in the title. The proposed technique performed well in terms of precision and f-score, outperforming the results obtained using NeFGIS. Sentence relevant to the title proposing a technique for improving the summary result using a word-based similarity approach and a meaning-based similarity approach. The word-based similarity compared two phrases based on the n-word sequence with n-gram, and the meaning-based similarity scored two sentences using query expansion with Thesaurus. The proposed technique has a high f-score based on two elements of similarity: word-based and meaning-based similarity. F-score is the harmonic mean of recall and precision; the highest possible value of an f-score, signifying flawless precision and recall.

Table 4 displays the average recall, precision, and f-score for all Rouge measures using the proposed approach versus the NeFGIS method. The NeFGIS technique outperforms the proposed method by 0.26065 on average recall. However, the proposed method has a better average evaluation result in precision and f-score when compared to NeFGIS, with sequentially being 0.04397 (11,59 percent increase) and 0.0084. (2,72 percent increase). The proposed method can improve the weighing of summaries in multi-documents using POS tagging and the sentence relevance to the title approach.

Testing was performed using the top ten sentences for the proposed method and NeFGIS. The sentences were compared using three different ground truths. Ground truth is obtained manually by experts by selecting representative sentences that represent a topic. Table 5 shows that the basic term in Indonesia can have multiple words and each word contains different POS label. The proposed method does not treat every word equally because it does not necessarily have the same function. POS labels will affect every term's weight in the sentence and mark a term with the most important term in the document. The sentence with the most content-bearing labels can have high sentence weight and can be sentence representative for summaries.

Table 4. Average evaluation result

Term	Recall	Precision	F-Score
Proposed method	0.25345	0.42319	0.31675
NeFGIS method	0.26065	0.37922	0.30835

Table 5. Example of basic terms with multiple POS labels

Basic term	POS	Example word	
		Indonesian word	English meaning
<i>laku</i>	Noun	<i>pelaku</i>	person/performer
	Verb	<i>melakukan</i>	do/perform
		<i>berlaku</i>	apply, be valid
		<i>memperlakukan</i>	treat
<i>tahu</i>	Adjective	<i>laku</i>	salable/saleable
	Noun	<i>pengetahuan</i>	knowledge
		<i>Tahu</i>	tofu
<i>tindak</i>	Verb	<i>mengetahui</i>	knowing/understand
		<i>tahu</i>	know
	Noun	<i>tindakan</i>	action
Verb	<i>penindakan</i>	prosecution	
	<i>menindak</i>	take action	

#### 4. CONCLUSION

This research looked at a sentence weighting method for multi-document summarization that selects meaningful sentences by combining news features with part of speech tagging and a sentence relevance approach. The proposed technique worked well and produced a better summary than the NeFGIS. The harmonic mean of recall and precision (f-score) for ROUGE-1, ROUGE-2, ROUGE-L, and ROUGE-SU4 news features with POS tagging and sentence relevance approach are 0.00717 (1.84 percent increase), 0.00662 (3.03 percent increase), 0.01481 (3.85 percent increase), and 0.00501 (3.85 percent increase) (2.08 percent increase). These are superior to the results produced using NeFGIS. The rising number is the result of calculating the weighting of words with sentence labels and the title's most relevant sentence. The results suggest that employing part of speech tagging and sentence relevance to the title, the proposed method can improve the weighting for news multi-document summarization in Bahasa Indonesia.

#### REFERENCES




- [1] A. Fanani, Y. Farida, P. P. Arhandi, M. M. Hidayat, A. Muhid, and B. Montolalu, "Regression model focused on query for multi documents summarization based on significance of the sentence position," *Telecommunication Computing Electronics and Control TELKOMNIKA*, vol. 17, no. 6, pp. 3050-3056, 2019, doi: 10.12928/TELKOMNIKA.v17i6.12494.
- [2] D. R. Radev, E. Hovy, and K. McKeown, "Introduction to the Special Issue on Summarization," *Computational Linguistics*, vol. 28, no. 4, pp. 399-408, 2002.
- [3] J. Atkinson and R. Munoz, "Rhetorics-based multi-document summarization," *Expert Systems with Applications*, vol. 40, no. 11, pp. 4346-4352, 2013, doi: 10.1016/j.eswa.2013.01.017.
- [4] T. He, F. Li, W. Shao, J. Chen and L. Ma, "A New Feature-Fusion Sentence Selecting Strategy for Query-Focused Multi-document Summarization," *International Conference on Advanced Language Processing and Web Information Technology*, 2008, pp. 81-86, doi: 10.1109/ALPIT.2008.45.
- [5] X. Wan, J. Yang, and J. Xiao, "Towards an Iterative Reinforcement Approach for Simultaneous Document Summarization and Keyword Extraction," *Annual Meeting of the Association of Computational Linguistics*, no. June, pp. 552-559, 2007.
- [6] M. Z. Abdullah and C. Fatichah, "Multi-document summarization based on news features and part of speech tagging (in Indonesia: Peringkasan multi-dokumen berita berdasarkan fitur berita dan part of speech tagging)," *Register: Jurnal Ilmiah Teknologi Sistem Informasi*, vol. 4, no. 2, pp. 56-66, 2018, doi: 10.26594/register.v4i2.1251.
- [7] S. Manne and S. S. Fatima, "A Feature Terms based Method for Improving Text Summarization with Supervised POS Tagging," *International Journal of Computer Applications*, vol. 47, no. 23, pp. 7-14, 2012.
- [8] K. Umam, F. W. Putro, G. Qorik, O. Pratamasunu, A. Z. Arifin, and D. Purwitasari, "Coverage, Diversity, And Coherence Optimization For Multi-Document Summarization," *Jurnal Ilmu Komputer dan Informasi*, vol. 8, no. 1, pp. 1-10, 2015, doi: 10.21609/jiki.v8i1.278.
- [9] A. Wahib, A. Z. Arifin, and D. Purwitasari, "Improving Multi-Document Summary Method Based on Sentence Distribution," *Telecommunication Computing Electronics and Control TELKOMNIKA*, vol. 14, no. 1, pp. 286-293, 2016, doi: 10.12928/TELKOMNIKA.v14i1.2330.
- [10] D. R. Radev, H. Jing, M. Styś, and D. Tam, "Centroid-based summarization of multiple documents," *Information Processing and Management*, vol. 40, no. 6, pp. 919-938, 2004, doi: 10.1016/j.ipm.2003.10.006.
- [11] R. Ferreira *et al.*, "Assessing sentence scoring techniques for extractive text summarization," *Expert Systems with Applications*, vol. 40, no. 14, pp. 5755-5764, 2013, doi: 10.1016/j.eswa.2013.04.023.
- [12] M. Fachrurrozi, N. Yusliani, and R. U. Yoanita, "Frequent Term based Text Summarization for Bahasa Indonesia," *International Conference on Innovations in Engineering and Technology*, 2013, pp. 30-32.
- [13] A. Z. Arifin, M. Z. Abdullah, and A. W. Rosyadi, "Sentence Extraction Based on Sentence Distribution and Part of Speech Tagging for Multi-document Summarization," *Telecommunication Computing Electronics and Control TELKOMNIKA*, vol. 16, no. 2, pp. 843-851, 2018, doi: 10.12928/TELKOMNIKA.v16i2.8431.
- [14] Abdi, N. Idris, R. M. Alguliyev, and R. M. Alguliyev, "Query-based multi-documents summarization using linguistic knowledge and content word expansion," *Soft Computing*, vol. 21, no. 7, pp. 1785-1801, 2017.
- [15] C. Blake and Coling, "A Comparison of Document, Sentence, and Term Event Spaces," *Coling/Acl 2006*, Vols 1 and 2, Proceedings of the Conference, pp. 601-608, 2006.
- [16] Y. K. Meena and D. Gopalani, "Evolutionary algorithms for extractive automatic text summarization," *Procedia Computer Science*, vol. 48, no. C, pp. 244-249, 2015, doi: 10.1016/j.procs.2015.04.177.
- [17] J. P. Mei and L. Chen, "SumCR: A new subtopic-based extractive approach for text summarization," *Knowledge and Information Systems*, vol. 31, no. 3, pp. 527-545, 2012.
- [18] C. Sunitha, A. Jaya, and A. Ganesh, "Automatic summarization of Malayalam documents using clause identification method," *International Journal of Electrical and Computer Engineering*, vol. 9, no. 6, pp. 4929-4938, 2019.
- [19] H. Sujaini, K. Kuspriyanto, A. Akhmad Arman, and A. Purwarianti, "A Novel Part-of-Speech Set Developing Method for Statistical Machine Translation," *Telecommunication Computing Electronics and Control TELKOMNIKA*, vol. 12, no. 3, pp. 581-588, 2014, doi: 10.12928/TELKOMNIKA.v12i3.79.
- [20] I. N. P. Trisna and A. Nurwidyantoro, "Single document keywords extraction in Bahasa Indonesia using phrase chunking," *Telecommunication Computing Electronics and Control TELKOMNIKA*, vol. 18, no. 4, pp. 1917-1925, 2020, doi: 10.12928/TELKOMNIKA.v18i4.14389.
- [21] C. Lioma and R. Blanco, "Part of Speech Based Term Weighting for Information Retrieval," *ECIR 2009: Advances in Information Retrieval*, 2009, pp. 412-423.
- [22] D. Jurafsky and J. H. Martin, "N-gram, Speech and Language Processing," *An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, pp. 29-54, 2014.
- [23] H. Imran and A. Sharan, "Thesaurus and Query Expansion," *International Journal of Computer science & Information Technology IJCSIT*, vol. 1, no. 2, pp. 89-97, 2009.
- [24] Á. F. Zazo, C. G. Figuerola, J. L. A. Berrocal, and E. Rodríguez, "Reformulation of queries using similarity thesauri," *Information Processing and Management*, vol. 41, no. 5, pp. 1163-1173, 2005, doi: 10.1016/j.ipm.2004.05.006.






- [25] C. Y. Lin, "Rouge: A package for automatic evaluation of summaries," *Proceedings of the workshop on text summarization branches out (WAS 2004)*, no. 1, pp. 25-26, 2004.

## BIOGRAPHIES OF AUTHORS



**Moch Zawaruddin Abdullah**    received the associate degree in Informatics Engineering from State Polytechnic of Bandung (POLBAN) in 2010, and the bachelor of Applied Science in Informatics Engineering from Electronic Engineering Polytechnic Institute of Surabaya (PENS) in 2013. Master of Computer in Informatics Engineering from Sepuluh Nopember Institute of Technology (ITS), Indonesia, in 2018. Currently, he is a lecturer at the Department of Information Technology, State Polytechnic of Malang (POLINEMA) and takes part in the Intelligent Systems and Information Systems Research Group. His research interests are information retrieval, data mining, text mining, artificial intelligent, decision support system, information system. He can be contacted at email: [zawaruddin@polinema.ac.id](mailto:zawaruddin@polinema.ac.id).



**Chastine Fatichah**    received the Ph.D. degree from the Tokyo Institute of Technology, Japan, in 2012. She is currently an Associate Professor with the Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia. She has published more than 110 journal articles and conference papers related to computer science. Her research interests include artificial intelligence, data mining, and image processing. She can be contacted at email: [chastine@if.its.ac.id](mailto:chastine@if.its.ac.id).