❒   2299

Enhancing Performance in Medical Articles Summarization with Multi-Feature Selection

Susetyo Bagas Bhaskoro, Saiful Akbar, Suhono Harso Supangkat
Institut Teknologi Bandung, Indonesia

ABSTRACT

The research aimed at providing an outcome summary of extraordinary events information for public health surveillance systems based on the extraction of online medical articles. The data set used is 7,346 pieces. Characteristics possessed by online medical articles include paragraphs that comprise more than one and the core location of the story or important sentences scattered at the beginning, middle and end of a paragraph. Therefore, this study conducted a summary by maintaining important phrases related to the information of extraordinary events scattered in every paragraph in the medical article online. The summary method used is maximal marginal relevance with an n-best value of 0.7. While the multi feature selection in question is the use of features to improve the performance of the summary system. The first feature selection is the use of title and statistic number of word and noun occurrence, and weighting tf-idf. In addition, other features are word level category in medical content patterns to identify important sentences of each paragraph in the online medical article. The important sentences defined in this study are classified into three categories: core sentence, explanatory sentence, and supporting sentence. The system test in this study was divided into two categories, such as extrinsic and intrinsic test. Extrinsic test is comparing the summary results of the decisions made by the experts with the output resulting from the system. While intrinsic test compared three n-Best weighting value method, feature selection combination, and combined feature selection combination with word level category in medical content. The extrinsic evaluation result was 72%. While intrinsic evaluation result of feature selection combination merger method with word category in medical content was 91,6% for precision, 92,6% for recall and f-measure was 92,2%.

Corresponding Author:

Susetyo Bagas Bhaskoro,
School of Electrical Engineering and Informatics,
Institut Teknologi Bandung,
Ganesha 10 Bandung, Indonesia.
Email: susetyo.bagas@s.itb.ac.id

1.   INTRODUCTION

The availability of medical information is always increasing, not only from medical records alone, but also from the community active participation. Participation is divided into two categories, such as, writing in the form of non-formal language that is sharing disease history experience and its recovery which are written to social media [1], [2]. Another category is formal writing that is usually written into online medical articles in the form of health demographic information and extraordinary events [3]-[5]. The health-related formal writing can be used as an alternative to automatic and rapid data collection for the needs of public health surveillance information, compared to the manual collection of reports from health care facility such as health centers, hospitals and clinics.

This study summarizes health information from online medical article. The number of available medical articles is plenty and has diverse information, so it can cause its own problems. Common problem that occurs from readers in closed domain such as medical is the speed of reading time and understanding the essence of the story of an online medical article [6], [7]. A summary of the text is used to make the information shorter while still maintaining important phrases contained in the medical article. The summary techniques explored in this study include determining the n-Best maximal marginal relevance (MMR) value, utilizing multi feature selection and weighting to improve the performance of the summary results.

Several studies pertinent to textual summaries have been made by some previous researchers. The techniques explored are differentiated into feature selection [7]-[14], weighting [15]-[17] and MMR. This is due to their simplicity, effectiveness and they yield relevant and non-exaggerated outputs [15], [18]-[21]. Vishal Gupta [9] used cue method, title, and location sentences as query or keyword. P. Y Zhang [14] stated that the select sentence used is similarity measure between sentences, word form similarity, word order similarity, word semantic similarity and sentence similarity. Dharmendra Hingu [12] explained that the feature selection that can be used for query or keyword includes relative position of sentence; named entities; similarities with other sentences; similarity with rest of the document; similarities with other sentences; title relevance; relative length of sentences; frequency of word; citation and numerical data. E. Padmalahari [7] and P. Goyal [11] used a combination of statistics and linguistics. Features used include acronym, keyword features, sentence position, term-frequency, length of the word, part of speech and proper noun feature, pronouns. Róbert Móro [13] explained that the paragraph initial location and the end of the paragraph have an important meaning, due to the information in that position has a positive value to be processed. Masanori Akiyama [10] mentioned that it takes the ranking of the summary results using jacquard coefficient. Vahdani [22], explains that unimportant sentences can be measured from the number of occurrences in the article. Researcher mentioned that frequent sentences can be obtained using word frequency calculation through the tf-idf method. However, the researcher did not mention the pre-processing stages used and did not mention additional techniques such as n-grams to reduce the calculation errors of the tf-idf method. So this research still has an open opportunity for improved evaluation result. Fauzi [23] offers proposed feature selection utilization using information gain and MMR as well as combines information gain and MMR. The obtained output shows that using a combined information gain and MMR yields 86%. Liu [24] conducted an exploration to get important information from the review result called "feature opinion" by using conditional random field method. Feature opinion proposes patterns in Chinese language and classifies positive and negative words.

In addition to feature selection, according to other researchers, weighting and n-Best are not the least important [15]-[17], [25], [26]. Reza Zaefarian utilized weighting tf-idf with intrinsic test results of 60%-70%. Gabriel Murray [16] and Sonia Haiduc [27] compared some weightings such as tf-idf, residual idf, tf, gain, and su-idf. Other studies explored merely on the use of document frequency (DF). The researchers said that DF can be used as feature selection to produce relevant information [28].

Several previous studies have described feature selection and feature suggestions to maintain important sentences in their summary results. However, from several studies that have been available, the selection of weight and n-Best value did not mentioned the best results. Therefore, this study will present the results of n-Best value exploration in the summary system. In addition, this research also explores multi-feature selection consisting of n-Best weighting value method, feature selection combination, and combined feature selection combination with word level category in medical content. Overall, this study aims to contribute as follows:
a. Generate the most appropriate n-Best value for the summary system in Indonesian medical articles;
b. Produce characteristic analysis for feature selection combination in summary system;
c. Provide a list of sentence patterns consisting of core sentences, explanatory sentences and supporting sentences.

The composition of writing in this study is presented as follows: The materials and methods were described in Section 2. In Section 3, described the result and analysis of the research. In Section 4, described the conclusion of the research.

2.   RESEARCH METHOD

The proposed system is shown in Figure 1. The used summary system utilized an extractive technique which is based on statistic or frequency. The purpose of applying extractive approach is to preserve messages conveyed by the author of the article.

Figure 1. Medical information extraction

Based on Figure 1, the research began with the Indonesian medical article extraction into a collection of sentences s (i) … s (n). Next is performing a test for feature selection, weighting and n-Best as well as classifying the word category level in medical content to prove that the summary result from the system is relevant to the summary result carried out manually. Moreover, Evaluations are divided into two categories: intrinsic evaluation; the system performance test, and extrinsic test; a test of an expert judgment. Table 1 is the characteristic of research text summary that became the reference.

Table 1. Characteristics of Research Text Summary

Properties	Characteristics
Feature selection	*title; noun; statistic number of word occurance; word range; statistic number of word and noun occurance; statistic number of word and title occurance*
Weighting	*Tf, Tf- idf, Tf-idf-df*
Parametrics value	0.4;0.6;0.7;0.8

2.1. Feature selection

Feature selection is an activity that specifies feature to serve as "query or keyword" that is used in summary system [12]. The following is pseudo-code for feature selection.

```
Pseudo Code  Feature Selection and Weighting Combination
1  : input            : document as d, feature_selection as fs weight as w
2  : output           : summary
3  :       query or keyword (□)={title; noun; statistic number of word occurance; word range; statistic number of word and noun
occurance; statistic number of word and title occurance}
4  :       weight (w)= {tf; tf-idf; tf-idf-df}
5  :       value parameter (□□□□□□□□□□□□□□□□□□□□□□□
6  :   di ← get a number of documents i
7  :   fs ← get value □ from the list
8  :   w ← get value w from the list
9:  combination of different queries ← get value from combination from fs and w
10: foreach (selection di is not null)
11:       sentence (si) = sentence detection from di
12:       word = word detection from si
13: end foreach
14: if(combination of different queries□ □is not null)
15:       si = each (si to sn) compared with □
16:       word = word detection from si  compared with □
17:       statistic = get statistic from (si to sn)
18:   w = tf (si to sn) x idf (si to sn)
19:     mmr (si) = π x Sim₁(si,d) − (1 − π)x Sim₂(si)
20: end if
21: summary ← if (si to sn) has bigger value than threshold, then (si to sn) is a   summary
```

2.1.1. Feature selection of title

It uses title feature to be query or keyword. Even though the result of initial study title cannot always be used to describe the content of the article, nonetheless, its reliability is able to produce a relevant

and appropriate summary. Where $N(t)$ is the number of words (t), while ρ is pre-processing, and k is the title keyword.

$$feature\ selection\ of\ title = N(t)/\rho(k) \tag{1}$$

2.1.2. Feature selection of noun
It uses noun feature to serve as query or keyword. The use of noun is due to meaning formation of the sentences are derived from collection of verbs or nouns. Where $N(t)$ is the number of words (t), while ρ is pre-processing, and t_i is words compared to list of words in (n). If $t_i \neq n$, then t_i is omitted.

$$feature\ selection\ of\ noun = \frac{N(t)}{\sum_{i=1}^{\infty} \rho(t_i \mid n)} \tag{2}$$

2.1.3. Feature selection statistic number of word occurence
It uses statistic number of word occurence feature to be query or keyword. The use of this feature is on the assumption that statistic number of word occurence is a conclusion from the core discussion in the article. Where $N(t)$ is the number of words (t), while ρ is pre-processing, and t_i is words compared to threshold t_h. If $t_h <$ max t_i, then ti is omitted.

$$feature\ selection\ statistic\ of\ word\ occurence = \frac{N(t)}{\sum_{i=1}^{\infty} \rho((t_i)>th)} \tag{3}$$

2.1.4. Feature selection word range
It uses word range feature to be a query or keyword. The use of this feature is on the assumption that word range is a unique query or keyword to describe the core discussion in the article. Where $N(t)$ is the number of words (t), while ρ is pre-processing, and t_i is words compared to max t_i.

$$feature\ selection\ word\ range = \frac{N(t)}{\sum_{i=1}^{\infty} \rho(max(t_i)>(t_i))} \tag{4}$$

2.1.5. Feature selection statistic number of word and noun occurrence
It uses statistic number of word and noun occurance feature to be query or keyword.

$$feature\ selection\ statistic\ number\ of\ word\ and\ noun\ occurance = \frac{N(t)}{\sum_{i=1}^{\infty} \rho(t_i \mid n)\ x\ (max(t_i)>(t_i))} \tag{5}$$

2.1.6. Feature selection statistic number of word and title occurrence
It uses statistic number of word and title occurance feature to be query or keyword.

$$feature\ selection\ statistic\ number\ of\ word\ and\ title\ occurance = \frac{N(t)}{\sum_{i=1}^{\infty} \vartheta(k)\ x\ (max(t_i)>(t_i))} \tag{6}$$

2.2. Weighting
In addition to feature selection, this study has also explored weighting. tf-idf weighting in the summary system has been used by many researchers [15], [24] and the obtained result is quite good. However, some researchers use only weighting tf. Tf weigth is used to calculate frequency of word occurrence from the entire document. The more the frequency of occurrence of the word, the higher the value of the weight. This study used mmr method for summary system as seen in the Equation (7).

$$mmr\ (Si) = \pi\ x\ Sim_1(Si,d) - (1-\pi)x\ Sim_2(Si,Sum) \tag{7}$$

Where d is an article in the vector form, and sum is collections of sentences extracted to be summary output. Sim1 and sim2 are used to calculate the similarity level from the article. Parametrics value variable is n-Best to balance the the summary with the most advisable output. N-Best that will be compared are 0.4, 0.6, 0.7, 0.8. Meanwhile, the similarity technique used is vector space model to compare two similar articles d_1, d_2. A technique to get query or keyword similarity with the content of the article is jaccard coefficient. Data set used in this research is as much as 7,346 pieces of medical article. The amount of data are obtained from two of the most popular sites in the health category including detik.com and kompas.com. Based on figure 2, the data set will be piloted and combined using feature selection and weighting. The number of combinations is 18 pairs.

$$f_k^n + w_k^n \approx f = \{x | x \le 6, x \in C\} + \{x | x \le 3, x \in C\} \qquad (8)$$

In addition to feature selection, another stage is to test the value of n-Best such as 0.4; 0.6; 0.7; 0.8.

2.3. Word level category in medical content

The next stage is to prove that the summary result generated by the system is in line with the result of summary completed by human manually. This verification uses the machine learning approach, and through this approach, the feature exploration will be seen as a classification problem. Furthermore, classification is done by dividing summary results into three categories of words in medical content. Word level category pattern in medical content is one of the feature selections used to find important phrases in online medical articles.

Table 2. Word Level Category in Medical Content

	Core sentences	Supporting sentences	Explanatory sentences
Pattern	Sentences→ [{description}, {symptom}, {disease}, {cause}, {effect}]	Sentence→ [({number}, {object}, {example}, {comparison}, {place}, {question sentence}, {quote})]	sentence→ [{citation}, {exclamations}, {solution}]

2.4. Evaluation

The evaluation that was conducted is divided into two categories; intrinsic evaluation which is classification test result for the word category in the medical content performed by the system using the multinomial naïve bayes method. Another test is an extrinsic evaluation which is the evaluation of test result on the conformity of the outputs from the system judged by the expert decision. Particularly for the extrinsic evaluation, expert has different backgrounds, such as: (EI) Biological; (E2) Informatics; (E3) Linguistic; and (E4) Humaniora.

The existence of the expert is divided into two functions. The first function is the expert serves as a classification maker for the word category level in medical content as in Table 2. The second function is the expert as the evaluator, i.e., the subjective assignment to the conformity of the summary result generated by the summary system. The evaluation parameter given by the expert for summary result are grouped into five categories: (a) Score 1 if the summary is not relevant; (b) Score 2 if the summary is less accepted; (c) Score 3 if the summary result is quite acceptable; (d) Score 4 if the summary result is accepted; and (e) Score 5 if the summary result is greatly accepted. Percentage value of the evaluation results as below: (1) 0%-19,99% is strongly disagree, (2) 20%-39,99% is disagree, (3) 40%-59,99% is border agree, (4) 60%-79,99% is agree, (5) 80%-100% is strongly agree.

3. RESULT AND ANALYSIS

This study extracted articles in the category of a coarse-grained approach analysis, therefore the dataset used derived from online medical news with particular topics was a remarkable occurrence. One example of news sources used in this study is shown in Figure 2. Based on Figure 2, the number of words in the article is amounted to 305, and the important sentence obtained manually and made new knowledge is amounted to 102. There are about 33% important information that must appear in the article to make new knowledge.

Some important sentences that can be used as new knowledge of the articles contained in Figure 2 include:

1. Head of Health Service of Temanggung Regency, Suparjo said data of diarrhea patients in Sigedong Village until this morning reached 64 people.
2. He said there was a dead victim from the outbreak case. The victim is 75 years old, besides diarrhea, he also suffers from hypertension.
3. However, he said it was allegedly because the water consumed by society and is currently still under the research.
4. Temanggung Health Office has established a post in the village which opens 24 hours.
5. He also socialized to the community to implement clean and healthy life.
6. In addition, chlorine dispersion is distributed in the spring and water reservoir to reduce the number of bacteria and germs.

Figure 2. Extraordinary events information from online medical articles[1]

The important sentence is not only generated from the ranking of word frequency that appears in the article, but from the calculation of the existing important words in each paragraph. The typical Indonesian article writing pattern is usually done from a general description at the beginning of the paragraph, followed by supporting sentences located in the middle of the story content of the article. The last discussion tells about the conclusion in the form of a solution. Each important sentence in each paragraph will have a connection to the other sentences in different paragraphs. There are several dependencies between explanatory and explained sentences or sentences that provide information on causes and sentences that explain the results. For example, the sentence contained in number 1 has a relationship with the sentence contained in number 2 (diarrhea patient - there is a dead victim from the outbreak case). Sentence number 1 also still has a relationship with the sentence contained in number 3 (diarrhea patient - allegedly because of the consumed water). Sentence in the number 1 still has a relationship with the sentence contained in number 4 (Sigedong Village - Establish Posko).

Therefore this study divides the discussion category in each article into three parts, as seen in Table 2. Each category in Table 2 provides an overview that the discussion in each paragraph consists of patterns of words that describe important sentences in the article.

3.1. Test on n-best and weighting value

The summary method used is the MMR with the explored n-Best value is 0.4; 0.6; 0.7; 0.8. Test result from the n-Best values includes: (1) the value of n-Best 0.4 gets a more concise summary, but works well only in articles that are less than 200 words. (2) The value of n-Best 0.6 gets irrelevant summary results, there is a lot of ambiguous information. (3) The value of n-Best 0.7 obtains a more acceptable and relevant summary result with manual summarizing activities. (4) The value of n-Best 0.8 result is irrelevant summary and there are many sentences that turned to be elusive. Table 3 and Figure 3 display the results of comparison of the use of n-Best value.
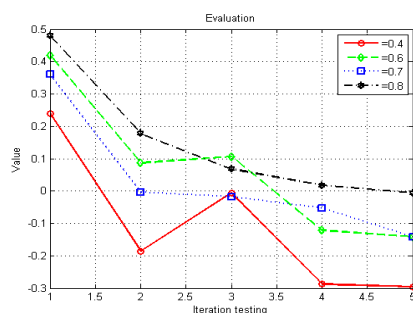

Figure 3. Graphic of n-Best Comparison

Table 3. Comparison of n-Best Value

n = 0,7	n = 0,8	n = 0,4	n = 0,6
0.36	0.48	0.42	0.24
-0.004	0.178	0.087	-0.186
-0.018	0.068	0.106	-0.006
-0.053	0.018	-0.122	-0.288
-0.142	-0.006	-0.142	-0.296
0.234	0.738	0.258	-0.536

Figure 3 shows a comparison graph of the n-Best value in a text summary study. Test result based on the utilization of weights shown in Table 4.

[1] https://lifestyle.okezone.com/read/2017/08/09/481/1752519/desa-sigedong-temanggung-klb-diare-1-korban-meninggal-dunia

Table 4. Result Comparison of Weighting Method

No	Properties	Tf-idf	Tf	Tf-idf-Df	Df
1	Presenting Basic order of the article	●		●	
2	Basic framework seems likely clear	●	●	●	●
3	Compressing main ideas into more concise one	●			
4	Presenting the article meaning	●			
5	Presenting supporting data	●	●	●	●
6	Presenting conclusion	●			
7	Summary results becomes less (20%)		●	●	●

Based on Table 4 above, if the summary results are fewer, then it becomes ambiguous, because the core sentence and supporting sentences are hard to come by. If the number of sentences from the summary results is almost the same as the original sentence, then the method in the system summary is not yet working. If the objective is to summarize in summary then the most appropriate weights are Tf-idf-df and df. The most appropriate thing in this research is Tf-idf.

3.2. Medical article extraction

The previous discussion shows that of the total number of words contained in the article, there are about 33% are important sentences. Therefore the first evaluation conducted in this research is to apply the appropriate feature selection and weighting to produce the word number output between 25 - 33%. Table 5 and Figure 4 are ten randomly drawn documents, and have different word counts. This preliminary test using the MMR with the value of n-Best 0.7 and get the result between 30%.
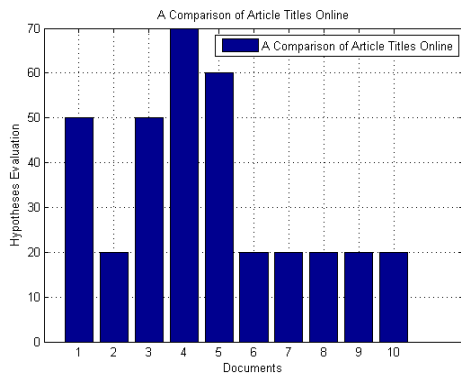

Figure 4. Comparisson of title with the content of the article

Table 5. Online Article Datasets and Dataset Evaluation

No	Posting	Extracted	Result
1	171	119	0.70
2	357	209	0.59
3	198	118	0.60
4	352	124	0.35
5	270	178	0.66
6	308	226	0.73
7	360	162	0.45
8	405	331	0.82
9	366	99	0.27
10	262	199	0.76

3.3. Test on feature selection combination

Combination of featured selection and weighting are as seen below.

1: $\{Title, Tf - Idf\}, \{Title, Tf\}, \{Title, Tf - Idf - Df\}$
2: $\{Noun, Tf - Idf\}, \{Noun, Tf\}, \{Noun, Tf - Idf - Df\}$
3: $\{Statistic, Tf - Idf\}, \{Statistic, Tf\}, \{Statistic, Tf - Idf - Df\}$
4: $\{LongWord, Tf - Idf\}, \{LongWord, Tf\}, \{LongWord, Tf - Idf - Df\}$
5: $\{NounStat, Tf - Idf\}, \{NounStat, Tf\}, \{NounStat, Tf - Idf - Df\}$
6: $\{TitleStat, Tf - Idf\}, \{TitleStat, Tf\}, \{TitleStat, Tf - Idf - Df\}$

Overall test results based on feature selection utilization is seen in table 6. Properties in Table 6 is the conclusion of the summary results obtained after testing using some combination of feature selection.

Based on Table 6, the feature selection results that have accurate output sequentially are the title feature, the statistical combination feature - noun, statistical feature, noun feature, statistical combination feature - the title and the longest word feature. Although the result of utilizing feature title gets a good ranking, but in some articles, the title feature may not be found and is very different from the content in the article. Completing the lack of the title feature utilization, this study will combine the title and statistics features from the number of noun occurrences. The comparison result of feature selection utilization and weighting to get more relevant summary result is as seen in Table 7.

☐ ISSN: 2088-8708

Table 6. Comparison Result of Feature Selection

	Properties	a	b	c	d	e	f
1	Relevant and suitable summary	•	•				•
2	Less suitable and ambiguous summary					•	
3	Irrelevant summary whose contents are mostly not suitable			•	•		
4	High level of accuracy			•			
5	Low level accuracy				•		

Note: (a) title; (b) statistic number of word occurance; (c) noun; (d) word range; (e) statistic number of word and title occurance; (f) statistic number of word and noun occurance;

Table 7. Comparison of Title with noun

TITLE + TFIDF					TITLE + TF					NOUN + TFIF					NOUN + TF				
1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
0.14	0.14	null	null	null	0.19	null	null	null	null	0.06	0.04	0.04	0.05	0.05	0.08	0.04	0.06	0.06	0.06
0.17	null	null	null	null	0.19	0.19	null	null	null	-	-	-	-	-	-	-	-	-	-
0.09	0.07	0.07	0.07	null	0.11	0.03	0.09	0.09	0.09	-	-	(0.01)	-	-	-	-	-	-	-
-	-	-	-	(0.03)	-	(0.02)	-	-	-	0.06	0.06	0.06	0.02	0.02	0.09	0.09	0.03	0.03	0.03
0.05	0.05	0.05	0.05	0.00	0.06	0.06	0.06	0.06	0.06	0.05	0.05	0.05	0.01	0.01	0.08	0.08	0.01	0.01	0.01
0.05	0.05	0.05	0.05	0.05	0.08	0.03	0.08	0.08	0.08	0.08	0.08	0.08	0.05	0.05	0.13	0.11	0.09	0.09	0.09
0.06	0.06	0.06	0.06	0.03	0.12	0.02	0.12	0.12	null	-	-	(0.01)	-	-	-	-	-	-	-
0.06	0.04	0.04	0.04	0.02	0.06	0.06	0.04	0.04	0.04	0.05	0.05	0.04	0.05	0.05	0.08	0.08	0.08	0.08	0.08
0.04	0.04	0.04	0.04	0.02	0.08	0.02	0.08	0.08	0.08	0.05	0.05	0.04	0.02	0.02	0.08	0.08	0.03	0.03	0.03

The test results give the following results 85.8% for precision, 83.7% for recall and f-measure is 84.7%.

3.4. Test on feature selection combination + word category level in medical content

This research combines feature selection with word category level in medical content. The objective is to keep important sentences while performing a summary. Merging such methods requires the sentence classification method in medical articles.

Pseudo Code Naïve Bayes Multinomial for The Classification of Sentence Structure
1: Calculate the naïve bayes multinomial to find the category of sentence from each test sentence by calculating the probability of each word type from the type of word found in the test sentence with each type of word in the training data sentence.
2: Looping based on test sentence
   a. Calculate the probability of each word type in the test sentence against the type forming each sentence category by using naïve bayes multinomial.
   b. Find the largest value of calculation output in each word type formation against the category of sentences under calculation.
   c. The formation of the word type against the category of sentence with the largest value is entered into the database. Data entered into are (sentence, set of word type on each word in sentence, sentence category).
   end

Table 8 shows the summary result by combining feature selection with word category level in medical content. Expected output is maintaining important sentences by following the patterns in each class of word category level in medical content.

Table 8. Result of Feature Selection Combination + Word Level Category Classificationin Medical Content

No	Medical Text	Classification Class
1	Polyphagia is one of three symptoms' diabetic disease	Explanatory sentences
2	Almost people doesn't aware about diabetic symptoms'	Supporting sentences
3	Someone must aware about some symptoms' like frequency to urination more often than before and always thirsty even they just drink, can be that diabetic symptoms'	Explanatory sentences
4	Alternative ways to prevent diabetic	Core sentences
5	According to diabetic international foundation at 2014, at least 70 percent from 9.1 million Indonesian people, realize she/he with diabetic after they got complicated disease	Supporting sentences
6	The easy way to detect diabetic with check your glucose blood regularly	Supporting sentences

No	Medical Text	Classification Class
7	People can prevent diabetic with healthy life style	Core sentences
8	Early detection to diabetic sign was important to keep you free from diabetic	Supporting sentences
9	Someone punished with diabetic when SGPT level higher than 126 mgdl and SGOT higher than 200 mgdl	Supporting sentences
10	Generally, people with diabetic should doing right and planned diet in low calories and fat	Supporting sentences
11	People with diabetic suggest to consume some food before hungry condition come because that condition can influence condition of their body	Explanatory sentences

Test result obtained by combining feature selection with word category level in medical content is as follows 91.6% for precision, 92.6% for recall and f-measure is 92.2%.

Table 9. The Calculation Output from System

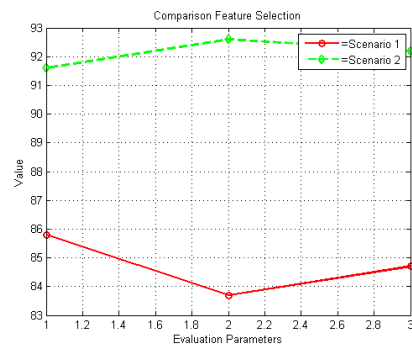| Supporting sentences	Explanatory sentences	Core sentences
5.76E-53	3.58E-55	1.20E-55
4.94E-38	8.28E-40	2.78E-40
3.73E-67	1.55E-70	5.23E-71
2.17E-49	7.39E-52	9.92E-52
2.47E-84	1.63E-89	9.85E-102
6.51E-96	1.45E-101	2.75E-90
1.65E-44	2.44E-47	1.64E-47

Figure 5. Comparison feature selection

3.5. Evaluation

Evaluation is divided into two categories, namely extrinsic and intrinsic. Extrinsic test involves an expert who has a role to assess a working system. The expert position in this research has several roles: (1) determining the sentence that must be produced in the summary from an article; (2) do the tagging sentences manually and (3) determine the word class between feature selections with word category level in medical content. The results obtained from the extrinsic evaluation were 72%. When viewed on the evaluation parameters associated with the percentage of the obtained value, then the decision is to agree.

Table 10. First Scenario of Extrinsic Supervised Test

a	b	c	d
6	13	5	1
7	7	4	3
9	21	8	1
6	11	6	0
6	14	6	0
5	11	5	0
11	8	8	3
4	5	3	1
9	10	7	2
5	7	4	1

Note: (a) Summary by The Expertise; (b) Summary by System; (c) Suitable Responses; (d) Non-Suitable Responses

Table 11. Second Scenario of Extrinsic Expert Opinion

Exp.	#1	#2	#3	#4
1	0.4	0.3	0.3	0.4
2	0.4	0.3	0.4	0.4
3	0.3	0.4	0.4	0.3
4	0.3	0.4	0.4	0.3
5	0.4	0.4	0.4	0.3
6	0.3	0.3	0.4	0.3
7	0.2	0.3	0.3	0.3
8	0.4	0.4	0.4	0.4
9	0.4	0.4	0.4	0.4
10	0.3	0.4	0.4	0.4

The second test is intrinsic. Table 9 shows the summary results generated from the system by combining feature selection with word category level in medical content. The conducted measurement consists of recall, precision and f-measure by comparing with other methods. Table 10 shows the first Scenario of Extrinsic Supervised Test. Table 11 shows the second Scenario of Extrinsic Expert Opinion. Table 12 shows the summary comparison results automatically. The comparable method is the DF feature selection as suggested by other researchers. In addition to the MMR-FS method used in Indonesian, Feature Selection Combination is defined as method one, Feature Selection Combination + Word Category Level in Medical Content defined as method two. Method one is a combination of feature selection that is also used

by some other researchers. In this study, we conducted a combination to get the most relevant summary results with manual summarizing activities. The method we proposed is method two by using multi-feature selection.

Table 12. The Comparison of Automatic Summarization

Methods	Recall	Precission	F-Measure
DF	95,868%	95,875%	95,871%
MMR-FS	-	86%	-
Method One	83,7%	85,8%	84,7%
Method Two	92,6%	91,6%	92,2%

When viewed from Table 12, the results of method two are still smaller by the DF method as what has been conducted by other researchers. However, the analysis results obtained in Table 4, the DF method can produce a more concise output taken from the summary. In other words DF is done for a fine-grained approach rather than in a coarse-grained approach.

4. CONCLUSION

Based on the research that has been conducted, it can be seen that every produced sentence must have at least one category of pattern in word category level in medical content. The result of intrinsic evaluation is 91,6% for precision, 92,6% for recall and f-measure is 92,2%. While extrinsic evaluation result is 72%. When viewed in the evaluation parameters related to the percentage of value, the final decision is conceded. Improved evaluation results can be done by adding techniques in the pre-processing stage.

REFERENCES

[1] S. Akbar, L. Slaughter, and Ø. Nytroø, "Collecting health related text from patient health writings", in *The 2nd International Conference on Computer and Automation Engineering*, 2010, vol. 1, pp. 15-19.

[2] A. Keselman, L. Slaughter, C. Arnott-Smith, H. Kim, G. Divita, A. Browne, C. Tsai, and Q. Zeng-Treitler, "Towards Consumer-Friendly PHRs: Patients' Experience with Reviewing Their Health Records", in *AMIA Annual Symposium Proceedings*, 2007, vol. 2007, no. February, pp. 399-403.

[3] M.H. Tekieh and B. Raahemi, "Importance of Data Mining in Healthcare : A Survey", pp. 1057-1062, 2015.

[4] S. Afantenos, V. Karkaletsis, and P. Stamatopoulos, "Summarization from medical documents: A survey", *Artif. Intell. Med.*, vol. 33, no. 2, pp. 157-177, 2005.

[5] C.D. Corley, D.J. Cook, A.R. Mikler, and K.P. Singh, "Text and structural data mining of influenza mentions in web and social media", *Int. J. Environ. Res. Public Health*, vol. 7, no. 2, pp. 596-615, 2010.

[6] N.B. and A. Jaiswal, "Literature Review on Automatic Text Summarization: Single and Multiple Summarizations", *Int. J. Comput. Appl.*, vol. 117, no. 6, pp. 20560-2948, 2015.

[7] E. Padmalahari, D.V.N.S. Kumar, and S. Prasad, "Automatic text summarization with statistical and linguistic features using successive thresholds", in *Proceedings of 2014 IEEE International Conference on Advanced Communication, Control and Computing Technologies, ICACCCT 2014*, 2014, pp. 1519-1524.

[8] Y. Li, W. Mao, D. Zeng, L. Huangfu, and C. Liu, "Extracting opinion explanations from Chinese online reviews", in *International Conference on Intelligence and Security Informatics: Cyberspace, Border, and Immigration Securities*, 2012, pp. 221-223.

[9] V. Gupta and G.S. Lehal, "A Survey of Text Summarization Extractive techniques", *J. Emerg. Technol. Web Intell.*, vol. 2, no. 3, pp. 258-268, 2010.

[10] M. Akiyama and K. Fujita, "How to Improve Patient Safety by Text Mining with Medical Incident Reports : Innovative Technologies Using e-Health and Health Technology Assessment", in *PICMET '13: Technology Management for Emerging Technologies*, 2013, pp. 2543-2550.

[11] P. Goyal, L. Behera, and T.M. McGinnity, "A context-based word indexing model for document summarization", *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 8, pp. 1693-1705, 2013.

[12] D. Hingu, D. Shah, and S.S. Udmale, "Automatic text summarization of Wikipedia articles", *Proc. - 2015 Int. Conf. Commun. Inf. Comput. Technol. ICCICT 2015*, pp. 15-18, 2015.

[13] M.B. Róbert Móro, "Personalized text summarization based on important terms identification", *Proc. - Int. Work. Database Expert Syst. Appl. DEXA*, pp. 131-135, 2012.

[14] P.Y. Zhang and C.H. Li, "Automatic text summarization based on sentences clustering and extraction", *Proc. - 2009 2nd IEEE Int. Conf. Comput. Sci. Inf. Technol. ICCSIT 2009*, vol. 1, no. 1, pp. 167-170, 2009.

[15] Y. Liu, S. Xie, and F. Liu, "Using N-best recognition output for extractive summarization and keyword extraction in meeting speech", *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, pp. 5310-5313, 2010.

[16] G. Murray and S. Renals, "Term-weighting for summarization of multi-party spoken dialogues", *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 4892 LNCS,

pp. 156-167, 2008.

[17] R. Zaefarian, "A New Algorithm for Term Weighting in Text Summarization Process", vol. 2006, pp. 292-297, 2006.

[18] C. Zhang, L. Zhang, C.J. Wang, and J.Y. Xie, "Text Summarization Based on Sentence Selection with Semantic Representation", *Proc. - Int. Conf. Tools with Artif. Intell. ICTAI*, vol. 2014–Decem, pp. 584-590, 2014.

[19] M.A. Mohamed and M. Oussalah, "Similarity-Based Query-Focused Multi-document Summarization Using Crowdsourced and Manually-built Lexical-Semantic Resources", *2015 IEEE Trust.*, no. June, pp. 80–-7, 2015.

[20] F. Boudin, M. El-b, and A. Cedex, "A Scalable MMR Approach to Sentence Scoring for Multi-Document Update Summarization", *Coling2008*, no. August, pp. 23-26, 2008.

[21] S. Xie and Y. Liu, "Using corpus and knowledge-based similarity measure in Maximum Marginal Relevance for meeting summarization", *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, no. 2, pp. 4985-4988, 2008.

[22] P.V. Amoli and O.S. Sh, "Scientific Documents ClusteringBased on Text Summarization", *Int. J. Electr. Comput. Eng.*, vol. 5, no. 4, p. 782-787, 2015.

[23] M.A. Fauzi, A.Z. Arifin, S.C. Gosaria, and I.S. Prabowo, "Indonesian News Classification Using Naïve Bayes and Two-Phase Feature Selection Model", vol. 8, no. 3, pp. 610-615, 2017.

[24] L.Z. Liu, W. Di Du, H.S. Wang, and W. Song, "Automatic Summarization in Chinese Product Reviews", *TELKOMNIKA (Telecommunication Comput. Electron. Control.*, vol. 15, no. 1, p. 373, 2017.

[25] A.M. and S.G. Niladri Chatterjee, "Single Document Extractive Text Summarization Using Genetic Algorithms", in *Third International Conference on Emerging Applications of Information Technology (EAIT) Single 2012*, 2012.

[26] A. Khan, N. Salim, and H. Farman, "Clustered Genetic Semantic Graph Approach for Multi-document Abstractive Summarization", in *Digital Information Processing and Communications (ICDIPC)*, 2016.

[27] S. Haiduc, J. Aponte, L. Moreno, and A. Marcus, "On the use of automated text summarization techniques for summarizing source code", *Proc. - Work. Conf. Reverse Eng. WCRE*, pp. 35-44, 2010.

[28] W. Zheng and G. Feng, "Feature Selection Method Based on Improved Document Frequency", *TELKOMNIKA (Telecommunication Comput. Electron. Control)*, vol. 12, no. 4, p. 905, 2014.

BIOGRAPHIES OF AUTHORS

Susetyo Bagas Bhaskoro received master degrees in informatics from School of Electrical Engineering and Informatics Engineering at the Institut Teknologi Bandung, in 2011. His research interest include medical application, biometrics recognition and data analytics.

Saiful Akbar received the bachelor and master degrees in Informatics from the Department of Informatics, Institut Teknologi Bandung, Indonesia, in 1997 and 2002, respectively, and the doctorate degree in Engineering Science from Johannes Kepler University of Linz, Austria, in 2007. He was also a Visiting Researcher at Norwegian University of Science and Technology (NTNU) in 2009-2010. He is currently an assistant professor at School of Electrical Engineering and Informatics, Institut Teknologi Bandung. His reaserch interests include data and knowledge engineering, multimedia database and similarity retrieval, information extraction, data analytics and visualization.

Suhono. H Supangkat is a Professor in the School of Electrical Engineering and Informatics Engineering at the Institut Teknologi Bandung, West Java, Indonesia. He received a master's degree from Meisei University, Japan, and a doctoral degree from the University of Electro-Communications, Tokyo, Japan, in 1998. His research interests include smart cities, smart education, smart health, smart energy, and smart mobility.