Jurnal Sains Komputer & Informatika (J-SAKTI)
Volume 7 Nomor 1, Maret 2023, pp. 450-461
ISSN: 2548-9771/EISSN: 2549-7200
https://tunasbangsa.ac.id/ejurnal/index.php/jsakti

Automatic Short Answer Assessment Using The Cosine Similarity Method

Samsu Bahri[1], Mugi Praseptiawan[2], Winda Yulita[3*]
[1]Program Studi Pendidikan Bahasa Inggris, STKIP Setia Budhi Rangkasbitung, Indonesia
[2,3]Program Studi Teknik Informatika, Institut Teknologi Sumatera, Indonesia
e-mail: [1] samsu.bahri@stkipsetiabudhi.ac.id, [2]mugi.praseptiawan@if.itera.ac.id,
[3]winda.yulita@if.itera.ac.id

Abstract

In the learning process, most exams to assess learning achievement have been carried out by providing questions in the form of short answers or essay questions. The variety of answers given by students makes a teacher have to focus on reading them. This process of assigning grades is difficult to guarantee quality if done manually. Moreover, each class is mastered by a different teacher, which can cause inequality in the grades obtained by students due to the influence of differences in teacher experience. Therefore the automated answer assessment research was developed. The automatic short answer assessment is designed to automatically assess and evaluate students' answers based on a trained set of answer documents. The automated grading system uses the cosine similarity method to determine the degree of similarity of a student's answer to the teacher's answer. While the word weighting used is the Term Frequency-Inverse Document Frequency (TF-IDF) method. The data used is a question totaling 5 questions with each question answered by 30 students, while the students' answers are assessed by experts to determine the real value. This study was evaluated by Mean Absolute Error (MAE) with the resulting value of 0.22.

Keywords: *auto rater, cosine similarity, TF-IDF, MAE*

1. INTRODUCTION

In the learning process, most exams to assess learning achievement have been carried out by providing questions in the form of short answers or essay questions. The variety of answers given by students makes a teacher have to focus on reading them. This process of assigning grades is difficult to guarantee quality if done manually. Moreover, each class is mastered by a different teacher can cause inequality in the grades obtained by students due to the influence of differences in teacher experience [1]. Therefore automated answer assessment research is developed.

Automatic short answer assessment is an important branch of intelligent education. The automatic short answer assessment is designed to automatically assess and evaluate students' answers based on a trained set of answer documents [2]. This system can help teachers check students' learning comprehension and reduce the teacher's influence in providing subjective grades. At the same time, automatic assessment can reduce labor costs and material resources. Therefore, the task of automated short answer assessment has significant convenience and great commercial value [3].

Automatic assessment has been widely researched and developed in various languages, such as Arabic, English, Indonesian and other languages. Automatic answer scoring in arabic is done by Abdeljaber [4] by using the Longest Common Subsequence (LCS) algorithm and using the Arabic WordNet language. In English it

Jurnal Sains Komputer & Informatika (J-SAKTI)
Volume 7 Nomor 1, Maret 2023, pp. 450-461
ISSN: 2548-9771/EISSN: 2549-7200
https://tunasbangsa.ac.id/ejurnal/index.php/jsakti

was developed by Chang using the machine-generated concept map method [5]. Development of automatic short answer scoring in Indonesian [6] done using similarity measurement (cosine similarity) to detect the similarity of students' answers with the teacher's answers. The study did not use the stage of measuring word weights before measuring the similarity of answers.

One of the methods commonly used to measure the weight of a word is the Term Frequency-Inverse Document Frequency (TF-IDF). The TF-IDF algorithm is a statistical measure used to evaluate how important a word is in one sentence in a text/document [7]. The TF-IDF method is a technique that is commonly known and used as a weighting technique and its performance is even still comparable to the new technique [8]. The TF-IDF method is widely applied in various fields, including information retrieval. The TF-IDF method is applied to information retrieval research to find news text that matches the given keywords. This research successfully displayed relevant or appropriate articles on the system created [9].

2. RESEARCH METHODOLOGY

Automatic short answer assessment is an important branch of intelligent education. The automatic short answer assessment is designed to automatically assess and evaluate students' answers based on a trained set of answer documents [2]. This system can help teachers check students' learning comprehension and reduce the teacher's influence in providing subjective grades. Figure 1 is a research flow that is carried out in stages, namely literature study, system design, system testing and analysis and discussion.

The literature study method is a series of activities related to the method of collecting library data, reading and recording, and managing research materials. The reference sources (books, journals, articles) referenced in this study are those related to the Automatic Exam Grading System for Essay Questions.

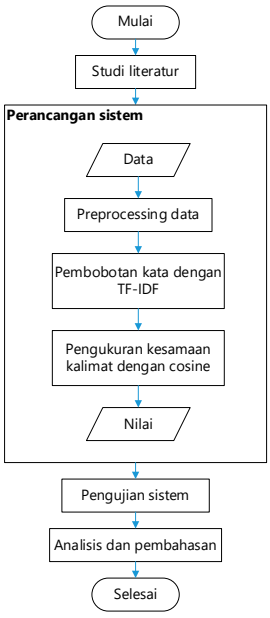

Figure 1. Research flow

The data used is the result of student essay answers totaling 5 people along with 5 essay questions and teacher answers from the 5 questions. The data also contains the scores obtained by students in each question with the value of each question, which is 20.

1. Preprocessing data

Text preprocessing is a stage to process news text that is raw material into words that are ready to be calculated the weight of the word. Some of the processes of text preprocessing, namely sentence segmentation, case folding, tokenizing, filtering, and stemming.

a) Segmentation of sentences
Sentence segmentation is the first step of the text preprocessing process. In this process, the news text consisting of paragraphs is broken into several sentences. The separation of each sentence is based on punctuation marks, such as a period (.), an exclamation mark (!) and a question mark (?).

b) Case folding
News paragraphs that have been cut into sentences will run the case folding process. Case folding is the process of converting all text into lowercase characters and discarding all characters other than a-z. If there are punctuation marks, numeric numbers and symbols are omitted.

c) Tokenizing
Tokenizing is a process to transform sentence forms into single words. The cutting of a sentence based on the delimiter that composes it, that is, a space (" ").

d) Filtering
In the filtering process, stopword removal is carried out. Stopwords are words that have no meaning or are less meaningful words and often appear in a collection of words.

e) Stemming
The next process is stemming, which is to return a word to its root form (root word) with certain rules, so that each word has the same representation. Stemming in this study using Nazief & Adriani algorithm [10].

2. Word weighting with TF-IDF

The TF-IDF (Term Frequency-Inverse Document Frequency) method is a way to give weight to the relationship of a word (term) to a document. This method combines two concepts for calculating weights, namely the frequency of occurrence of a word in a particular document and the inverse frequency of the document containing the word. The frequency with which a word appears in the provided document indicates how important the word is in the document. The frequency of the document containing the word indicates how common the word is. So the weight of the relationship between a word and a document will be high if the frequency of the word in the document is high and the overall frequency of the document containing the word in the document is low [11]. The word weighting formula of the TF-IDF modification is :

Jurnal Sains Komputer & Informatika (J-SAKTI)
Volume 7 Nomor 1, Maret 2023, pp. 450-461
ISSN: 2548-9771/EISSN: 2549-7200
https://tunasbangsa.ac.id/ejurnal/index.php/jsakti

$$w_{i,j} = \left(tf_{i,j} \, x \, log \left(\frac{N}{df_i} \right) \right) \quad (1)$$

Where $w_{i,j}$ is the weight of the word ke-i from the document ke-j. $tf_{i,j}$ is the term frequency, is the word count of the word ke-i in the document ke-j. $log \left(\frac{N}{df_i} \right)$ is the Inverse Document Frequency formula (IDF), N represents the sum of the entire document or sentence.

3. Sentence similarity gauge with cosine similarity

Cosine similarity is used to calculate the relevance/similarity of the teacher's answer to the student's answer. The relevance value is obtained by measuring the similarity between 2 vectors, namely the teacher's answer vector and the student's answer vector. The greater the relevance value, the more relevant the teacher's answer and the student's answer will be. According to Patidar et al. [12], a similarity measure is the distance between various data points. Similarity measure is also used in measuring the similarity between sets based on the intersection of two sets. Similarity measure is also known as a function that calculates the degree of similarity between a pair of text objects. In short, similarity is the amount that reflects the strength of the relationship between two data. One of the most commonly used measures of similarity is cosine similarity. Cosine similarity is the basis for calculations to obtain the value of relevance between queries and documents and relevance between documents. Cosine similarity is the cosine of the angle between vectors. Cosine similarity has a formula as belowθ:

$$sim\ (S_1, S_2) = \frac{\sum_i t_{1i} t_{2i}}{\sqrt{\sum_i t_{1i}^2} \, x \, \sqrt{\sum_i t_{2i}^2}} \quad (2)$$

Information :
S_1 = the weight vector of the word that is the candidate
S_2 = vectors of word weights other than candidates
Where t_i is the word weight of the word w_i.

4. Value

Grades are the result of system processing by measuring the similarity of teacher and student answers to each question.

5. System testing

Testing is carried out by looking at the student scores generated by the system with the grades given by the teacher manually. At this stage, an accuracy assessment of the value generated by the system will be carried out. To evaluate the performance of the system that has been made, a Mean Absolute Error (MAE) calculation will be used between the results of the system assessment and the results of manual assessment[13]. The equation used to calculate the correlation value is as follows:

$$MAE = \frac{1}{n} \sum_{i-1}^{n} |x_i - j_i| \quad (3)$$

Jurnal Sains Komputer & Informatika (J-SAKTI)
Volume 7 Nomor 1, Maret 2023, pp. 450-461
ISSN: 2548-9771/EISSN: 2549-7200
https://tunasbangsa.ac.id/ejurnal/index.php/jsakti

This stage discusses the results obtained for each step of the system design. Starting from the process of text preprocessing, namely sentence segmentation, case folding, tokenizing, filtering, and stemming to the application of measuring sentence similarity with the cosine similarity method. In addition, at this stage, it will also be discussed regarding the accuracy obtained by the system based on the values generated by the automated scoring system.

3. Results And Discussion

A. Preprocessing

The purpose of the text preprocessing stage is to convert news articles into words that are ready to be processed for word weight calculations. Some of the processes of text preprocessing, namely sentence segmentation, case folding, tokenizing, filtering, and stemming. The following is one example of a document that is inputted in the text preprocessing process accompanied by the stages of the text preprocessing process :

```
Question: What is Kevin Spacey's attitude in civil court?

Answer: Kevin Spacey testifies in his defense in civil trial. Kevin
Spacey has taken the stand as the first witness in his own defense in
the sexual misconduct trial against him, brought by actor Anthony
Rapp. In a response to the first question from his attorney, Chase
Scolnick, Spacey said Rapp's allegations are not true.
```

1) Sentence segmentation

It is the breaking of paragraphs into sentences. Solving is done by separating by separating by period punctuation (.), question mark (?) and exclamation mark (!). The results of the sentence segmentation process are shown in table 1

Table 1. Sentence segmentation results

No.	Kalimat
0.	Kevin Spacey testifies in his defense in civil trial
1.	Kevin Spacey has taken the stand as the first witness in his own defense in the sexual misconduct trial against him, brought by actor Anthony Rapp.
2.	In a response to the first question from his attorney, Chase Scolnick, Spacey said Rapp's allegations are not true.

2) Case folding

News paragraphs that have been cut into sentences will run the case folding process. Case folding is to convert all text to lowercase characters and discard all characters other than a-z. In addition, punctuation, numerical numbers and symbols are also omitted. Table 2 is the result of the case folding process.

Table 2. Case folding results

No.	Kalimat
0.	kevin spacey testifies in his defense in civil trial
1.	kevin spacey has taken the stand as the first witness in his own defense in

Jurnal Sains Komputer & Informatika (J-SAKTI)
Volume 7 Nomor 1, Maret 2023, pp. 450-461
ISSN: 2548-9771/EISSN: 2549-7200
https://tunasbangsa.ac.id/ejurnal/index.php/jsakti

No.	Kalimat
	the sexual misconduct trial against him brought by actor anthony rapp
2.	in a response to the first question from his attorney chase scolnick spacey said rapps allegations are not true

3) Tokenizing the word

It is the process of cutting sentences into words. Sentence cutting based on the delimiter that composes it, that is, the space (" ").

Table 3. Case folding results

No.	Word	No.	Word
1	kevin	28	trial
2	spacey	29	against
3	testifies	30	him
4	in	31	brought
5	his	32	by
6	defense	33	actor
7	in	34	anthony
8	civil	35	rapp
9	trial	36	in
10	kevin	37	a
11	spacey	38	response
12	has	39	to
13	taken	40	the
14	the	41	first
15	stand	42	question
16	as	43	from
17	the	44	his
18	first	45	attorney
19	witness	46	chase
20	in	47	scolnick
21	his	48	spacey
22	own	49	said
23	defense	50	rapps
24	in	51	allegations
25	the	52	are
26	sexual	53	not
27	misconduct	54	true

4) Filtering

In this filtering stage performs stopword removal. Stopwords are words that have no meaning or are less meaningful words and often appear in a collection of words. How to get rid of unimportant words by checking on the stopword dictionary. If the word exists the same as the stopword, then the word will be discarded or deleted.

Table 4. Hasil filtering

Jurnal Sains Komputer & Informatika (J-SAKTI)
Volume 7 Nomor 1, Maret 2023, pp. 450-461
ISSN: 2548-9771/EISSN: 2549-7200
https://tunasbangsa.ac.id/ejurnal/index.php/jsakti

No.	Kata	No.	Kata
1	kevin	18	against
2	spacey	19	brought
3	testifies	20	actor
4	defense	21	anthony
5	civil	22	rapp
6	trial	23	response
7	kevin	24	first
8	spacey	25	question
9	taken	26	attorney
10	stand	27	chase
11	first	28	scolnick
12	witness	29	spacey
13	own	30	said
14	defense	31	rapps
15	sexual	39	allegations
16	misconduct	40	true
17	trial		

5) Stemming

Stemming, that is, returning a word to its root form (root word), so that each word has the same representation. In this method it only handles affixes (affixes) prefixes (prefixes) and suffixes (suffixes) only. This is due to the rare occurrence of cases of addition of infix (insert) affixes in the Indonesian. The stemming results are shown in Table 5.

Table 5. Stemming results

No.	Kata	No.	Kata
1	kevin	18	against
2	spacey	19	brought
3	testifi	20	actor
4	defens	21	anthoni
5	civil	22	rapp
6	trial	23	respons
7	kevin	24	first
8	spacey	25	question
9	taken	26	attorney
10	stand	27	chase
11	first	28	scolnick
12	wit	29	spacey
13	own	30	said
14	defens	31	rapp
15	sexual	39	alleg
16	misconduct	40	true
17	trial		

B. TF-IDF

Jurnal Sains Komputer & Informatika (J-SAKTI)
Volume 7 Nomor 1, Maret 2023, pp. 450-461
ISSN: 2548-9771/EISSN: 2549-7200
https://tunasbangsa.ac.id/ejurnal/index.php/jsakti

After the text preprocessing process, the next stage is the calculation of word weights with the TF-IDF algorithm using words generated from the stemming process.

C. Cosine similarity

If the word weight has been obtained, it further looks for the cosine similarity value generated by the system. Table 6 is the value of the similarity between the answers of teachers and students produced by the system in question 1 is the same as question 5 (Q1-Q5).

Table 6. The similarity value of the system

No	Q1	Q2	Q3	Q4	Q5
1	0,67	0,56	0,74	0,68	0,56
2	0,73	0,78	0,75	0,58	0,67
3	0,34	0,65	0,93	0,56	0,54
4	0,52	0,77	0,71	0,85	0,76
5	0,36	0,68	0,38	0,81	0,98
6	0,56	0,74	0,58	0,59	0,45
7	0,78	0,56	0,59	0,63	0,56
8	0,89	0,56	0,93	0,77	0,56
9	0,45	0,78	0,84	0,85	0,39
10	0,52	0,94	0,58	0,74	0,78
11	0,53	0,85	0,83	0,83	0,56
12	0,69	0,86	0,85	0,59	0,78
13	0,88	0,88	0,83	0,66	0,56
14	0,83	0,93	0,77	0,59	0,87
15	0,75	0,75	0,71	0,84	0,67
16	0,56	0,78	0,59	0,72	0,53
17	0,72	0,77	0,63	0,82	0,69
18	0,93	0,73	0,68	0,56	0,78
19	0,47	0,76	0,92	0,56	0,89
20	0,23	0,64	0,47	0,67	0,63
21	0,56	0,73	0,57	0,79	0,78
22	0,45	0,82	0,65	0,58	0,73
23	0,69	0,57	0,56	0,37	0,56
24	0,83	0,59	0,47	0,58	0,78
25	0,87	0,73	0,78	0,67	0,67
26	0,79	0,67	0,59	0,68	0,67
27	0,84	0,94	0,61	0,49	0,52
28	0,59	0,45	0,77	0,79	0,65
29	0,79	0,83	0,74	0,56	0,43
30	0,88	0,82	0,69	0,73	0,89

The similarity value produced by the system ranges from 0 to 1. A similarity value close to the number 1 means that the similarity of answers between students

Jurnal Sains Komputer & Informatika (J-SAKTI)
Volume 7 Nomor 1, Maret 2023, pp. 450-461
ISSN: 2548-9771/EISSN: 2549-7200
https://tunasbangsa.ac.id/ejurnal/index.php/jsakti

and teachers is getting more and more similar, on the contrary, grades close to 0 state that students' and teachers' answers are increasingly not similar. The value obtained by the system is multiplied by the original score for each number, which is 20 points, so that the value generated by the system is between 1 and 20 for each number. The number of students is as many as 30 students.

Table 7. Conversion results in student 1

Dokumen	Nilai kemiripan sistem	Nilai expert
Q1	0,67 x 20 = 13,4	15
Q2	0,56 x 20 = 11,2	15
Q3	0,74 x 20 = 14,8	15
Q4	0,68 x 20 = 13,6	15
D5	0,56 x 20 = 11,2	10

Figure 1 to Figure 5 shows the comparison of the similarity value generated by the system with the value obtained from the expert for each question (Q1-Q5).

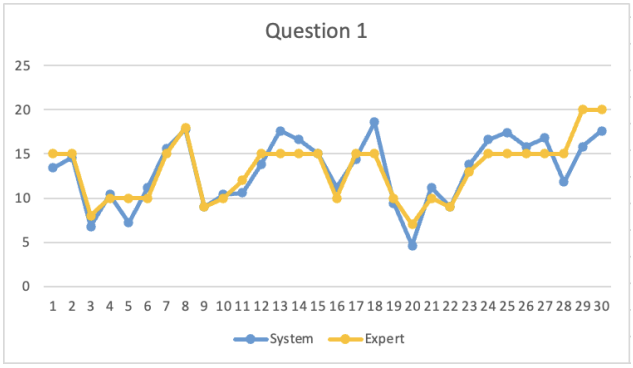

Figure 1. Comparison of values generated by systems & experts in Question 1

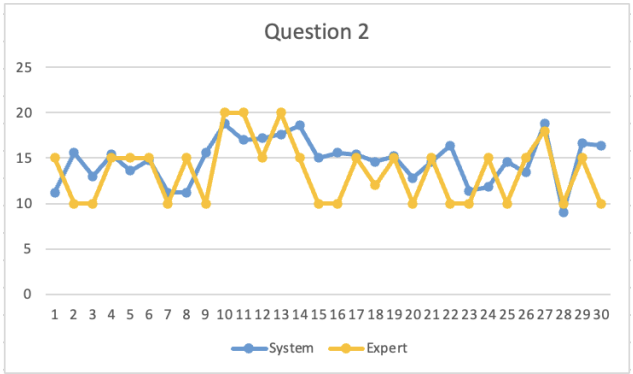

Figure 2. Comparison of values generated by systems & experts in Question 2

Jurnal Sains Komputer & Informatika (J-SAKTI)
Volume 7 Nomor 1, Maret 2023, pp. 450-461
ISSN: 2548-9771/EISSN: 2549-7200
https://tunasbangsa.ac.id/ejurnal/index.php/jsakti

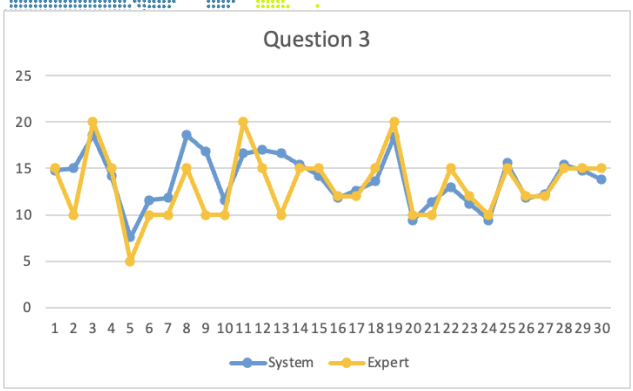

Figure 3. Comparison of values generated by the system & experts on the Question 3

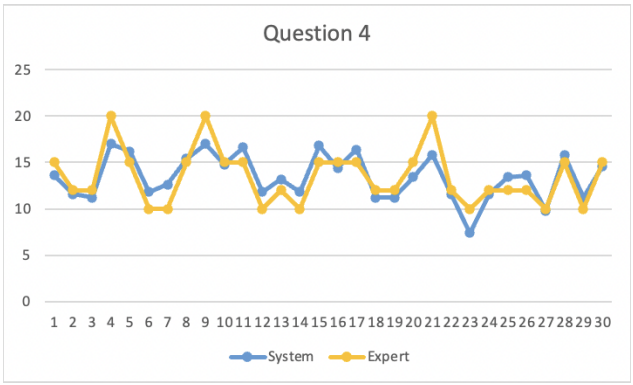

Figure 4. Comparison of values generated by systems & experts in Question 4

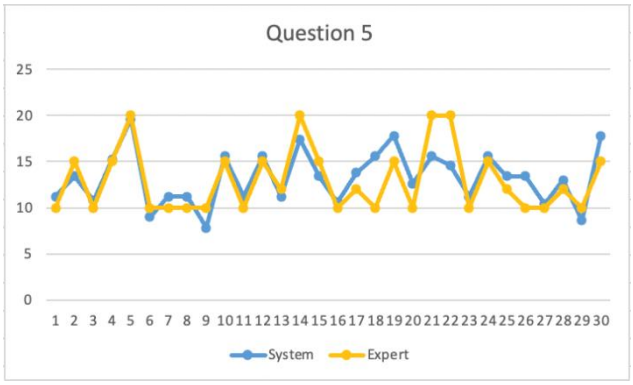

Figure 5. Comparison of values generated by systems & experts in Question 5

The application of the MAE (Mean Absolute Error) formula to the similarity value generated by the system with the value generated by the expert produces an MAE value of 0.02.

4. CONCLUSION

This research has successfully developed an automatic exam grading system for essay questions using the *cosine similarity* method. The system is evaluated

through the MAE (Mean Absolute Error) testing process with an error value of 0.22. Based on the results of the study conducted using data from a sample of student answers totaling 30 people with a total of 5 questions. The value generated by the system is multiplied by the actual value for each question. In this study, each similarity value of the system will be multiplied by 20 points. It will then be evaluated by using MAE to see the resulting error rate.

ACKNOWLEDGEMENT

This research was supported by Ministry of Education, Culture, Research, and Technology (Beginner Lecturer Research (PDP), Number : SP DIPA-023.17.1.690523/2022 tanggal 16 Juni 2022)

REFERENCES

[1] H. Sun and Z. Shao, "Research on the Application of Students' Answered Record Analyze Model and Question Automatic Classify Based on K-Means Clustering Algorithm," 2019 10th International Conference on Information Technology in Medicine and Education (ITME), 2019, pp. 494-497, doi: 10.1109/ITME.2019.00116.

[2] Z. Ke and V. Ng, "Automated essay scoring: A survey of the state of the art", Proc. 28th Int. Joint Conf. Artif. Intell., pp. 6300-6308, Aug. 2019.

[3] G. Lv, W. Song, M. Cheng and L. Liu, "Exploring the Effectiveness of Question for Neural Short Answer Scoring System," 2021 IEEE 11th International Conference on Electronics Information and Emergency Communication (ICEIEC)2021 IEEE 11th International Conference on Electronics Information and Emergency Communication (ICEIEC), 2021, pp. 1-4, doi: 10.1109/ICEIEC51955.2021.9463814.

[4] H. A. Abdeljaber, "Automatic Arabic Short Answers Scoring Using Longest Common Subsequence and Arabic WordNet," in IEEE Access, vol. 9, pp. 76433-76445, 2021, doi: 10.1109/ACCESS.2021.3082408.

[5] T. -H. Chang, J. -L. Chen, H. -M. Chou, M. -H. Bai, F. -Y. Hsu and Y. -C. Chen, "Automatic Scoring Method of Short-Answer Questions in the Context of Low-Resource Corpora," 2021 International Conference on Asian Language Processing (IALP), 2021, pp. 25-29, doi: 10.1109/IALP54817.2021.9675160.

[6] U. Hasanah, T. Astuti, R. Wahyudi, Z. Rifai and R. A. Pambudi, "An Experimental Study of Text Preprocessing Techniques for Automatic Short Answer Grading in Indonesian," 2018 3rd International Conference on Information Technology, Information System and Electrical Engineering (ICITISEE), 2018, pp. 230-234, doi: 10.1109/ICITISEE.2018.8720957.

[7] A. Mishra and S. Vishwakarma, "Analysis of TF-IDF Model and its Variant for Document Retrieval," 2015 International Conference on Computational Intelligence and Communication Networks (CICN), 2015, pp. 772-776, doi: 10.1109/CICN.2015.157

[8] M. Ramya and J. A. Pinakas, "Different type of feature selection for text classification", *International Journal of Computer Trends and Technology*, vol. 10, pp. 102-107, 2014

[9] A. N. Khusna and I. Agustina, "Implementation of Information Retrieval Using Tf-Idf Weighting Method On Detik.Com's Website," 2018 12th International Conference on Telecommunication Systems, Services, and Applications (TSSA), 2018, pp. 1-4, doi: 10.1109/TSSA.2018.8708744

[10] Nazief, B. A. A. dan M. Adriani. 1996. Confix-Stripping : Approach to Stemming Algorithm for Bahasa Indonesia. International Conference on Information and Knowledge Management, : 560-563

[11] Stephen Robertson, "Understanding Inverse Document Frequency: On Theoretical Arguments for IDF", *England Journal of Documentation*, vol. 60, pp. 502-520, 2005.

[12] Patidar, A. K., J. Agrawal dan N. Mishra. 2012. Analysis of Different Similarity Measure Functions and Their Impacts on Shared Nearest Neighbor Clustering Approach. International Journal of Computer Applications. 40(16): 1-5.

[13] W. Wang and Y. Lu, "Analysis of the Mean Absolute Error (MAE) and the Root Mean Square Error (RMSE) in Assessing Rounding Model", IOP Conference Series: Materials Science and Engineering, 2018.