Word Expansion using Synonyms in Indonesian Short Essay Auto Scoring

1st Nurul Chamidah
Faculty of Computer Science
Universitas Pembangunan Nasional Veteran Jakarta
Jakarta, Indonesia
nurul.chamidah@upnvj.ac.id

2nd Mayanda Mega Santoni
Faculty of Computer Science
Universitas Pembangunan Nasional Veteran Jakarta
Jakarta, Indonesia
megasantoni@upnvj.ac.id

3rd Helena Nurramdhani Irmanda
Faculty of Computer Science
Universitas Pembangunan Nasional Veteran Jakarta
Jakarta, Indonesia
helenairmanda@upnvj.ac.id

4th Ria Astriratma
Faculty of Computer Science
Universitas Pembangunan Nasional Veteran Jakarta
Jakarta, Indonesia
astriratma@upnvj.ac.id

5th Lomo Mula Tua
Faculty of Computer Science
Universitas Pembangunan Nasional Veteran Jakarta
Jakarta, Indonesia
lomomt@upnvj.ac.id

6th Trihastuti Yuniati
Fakultas Informatika
Institut Teknologi Telkom Purwokerto
Banyumas, Indonesia
trihastuti@ittelkom-pwt.ac.id

***Abstract*— Exams conducted in online learning to evaluate learning processes have many formats, including essay format. Essays are considered more proper to measure learning activity results. However, essays require longer to assess student answers and have consistency problems if the assessment is carried out by different teachers or done separately. This study investigates the influence of word expansion using synonyms in Indonesian thesaurus on short essay auto scoring. The first step, reference answers and student answer text data is preprocessed by case folding, stemming, stop word removal, tokenizing, and duplicate word removal. Second, Word expansion using synonyms in thesaurus is used to generate alternate words for reference answers. Third step, the scoring process by calculating similarity and matching words. The score from the similarity and matching results is then used to generate the final score. Performance evaluation shows that the Dice Coefficient similarity method achieved the highest correlation by a very good correlation, and the smallest MAE was achieved by the Cosine Coefficient similarity method.**

Keywords—essay auto scoring, word expansion, synonym, thesaurus

I. Introduction

Online learning or e-learning is currently applied in learning activities to deliver learning materials, assignments, and exams. Exams conducted to evaluate learning processes have many formats in online learning. Questions format commonly used to assess a lesson generally consists of two general formats, objective questions, and essay questions. Objective questions have answer choices that students can choose, such as multiple choice answers, matching answers, and true-false answers. Meanwhile, essay questions do not have answer choices. Students need to write their sentences to answer this question format. For example, write a free essay that does not use an answer key, and usually, teachers need to pay more attention to its grammar. There is also an essay question model that asks students to explain or describe something where the explanation has reference answers as the key answers.

Essay questions are considered more useful to estimate the depth of knowledge received by students from teaching [1] than objective questions that are easy to apply but difficult to measure the level of understanding. Although essays are considered more appropriate to measure the results of learning activities, this form of question requires a longer time to assess answers than evaluation of objective questions. Another problem assessing essay questions is the problem of consistency in both the same and different evaluators. Evaluation of essay questions by one person, inconsistency can occur if the review is carried out separately. Moreover, if other teachers carry out the assessment, they can get different scores on the same answer. Based on this explanation, it is necessary to create an auto scoring system to automatically assess essays to be more consistent.

Studies on automatic essay scoring in Indonesian have been done before. However, according to [2], there is not yet the most suitable method to use because automatic scoring system must provide an answer score as close as possible to the score given by human evaluators. The other problems, the resources for Indonesian are not as many as the resources in English. Studies [3]–[6] uses Latent Semantic Analysis (LSA) which maps keywords from student answers with keywords from reference answers into a matrix. Another study uses the Vector Space Model technique by word vectorizing using TF-IDF (Term Frequency-Inverse Document Frequency). It calculates cosine similarity between student answers and reference answers by taking keywords from these answers [7]. Study [8] utilizes lexical features in Natural Language Processing (NLP) by finding synonyms of words and then comparing between reference answers and student answers.

Scoring techniques for automatic essay scoring are also very diverse. In general, the assessment technique calculates closeness or similarity between the reference answer and student answer. The scoring method based on the similarity of keywords in a study [9] compares the scoring technique with the Longest Common Subsequence (LCS), Cosine Coefficient (CC), Jaccard Coefficient (JC), and Dice Coefficient (DC) in Indonesian short essays, also [10] comparing the last three methods. Likewise, some studies compare Cosine Similarity, Euclidean Distance, and Jaccard Coefficient [11]. Research [12] used K-Nearest Neighbour (KNN) to calculate the closeness between reference and student answers. Manhattan distance and dice similarity were used in the study [11] to calculate similarity answers and reference answers.

Word expansion can improve accuracy in text-based classification [13]. Word expansion using Indonesian thesaurus in the question answering system also shows better results than without using the query expansion [14]. Query expansion using Malay thesaurus also improve retrieval effectiveness[15].

This study aims to investigate the influence of word expansion using synonyms in Indonesian thesaurus to short essay auto scoring. This expansion applied to reference answers as alternative words to enrich terms. These enriching terms are carried out to increases the chance of similarity between reference answers and student answers.

II. METHODOLOGY

Fig.1 shows the method used in this study. The dataset contains text and numeric data types. Text data types consist of reference answers and student answers, and numeric data types consist of three scores for each student answer given by teachers. Preprocessing was conducted to reference answers and student answers. Word expansion using synonyms from Indonesian thesaurus is undertaken to expand preprocessed reference answers. The next step is calculating the similarity score and keyword matching score to obtain the final score. Finally, this study carries out performance evaluation to measure agreement between score given by teacher and score generated by the system.

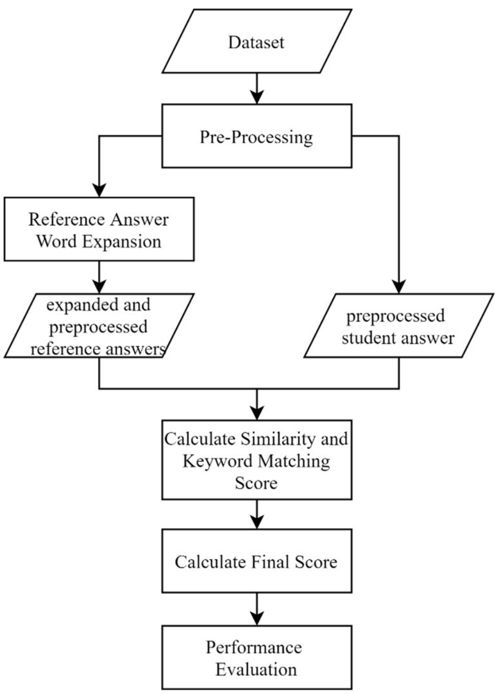

Fig. 1. Methodology

A. Dataset

The dataset used in this study consisted of 40 questions and reference's answer texts for related questions and 2162 answers obtained and used in previous studies [2], [16]. These 40 questions are divided into four categories they are politics, lifestyle, sport, and technology. The final numeric score for each student's answer is in the range [0, 100] that three teachers give. The data sample can be seen in TABLE I.

TABLE I. DATA SAMPLE

Feature	Value
Question	Apa yang dimaksud dengan volatile memory?
Reference Answer	Volatile memory adalah memory yang datanya dapat ditulis dan dihapus, tetapi hilang saat kehilangan power (kondisi off atau mati lampu).
Student Answer	Volatile merupakan memory yang datanya dapat ditulis serta dihapus, tetapi akan hilang jika tidak ada aliran listrik (kondisi mati/off) dan membutuhkan daya untuk mempertahankan memory.
Score 1	88
Score 2	100
Score 3	90

B. Preprocessing

Text preprocessing is done on student's answers and reference's answers. This process consists of several steps. They are case folding, stemming, stop word removing, tokenizing, and duplicate word removal. TABLE II. shows preprocessing from students answer and reference answer from TABLE I. Preprocessing stage which will be explained as follows:

- Case folding is a process to convert all characters into lowercase letters. This process includes removing all character's non-alphabet and punctuation. For example, the word "Volatile" became "volatile".
- Stemming is a technique that converts the word into root words. It removes prefixes, suffixes, and affixes. Sastrawi library based on Nazief-Adriani algorithm [17] is used to conduct this process. For example, words "datanya", "ditulis", "dihapus", etc. After stemming process, words became root words "data", "tulis", "hapus", etc.
- Stop word removing is used to remove words that have no meaning and usually have high occurrences. In this study, Sastrawi library stop word list is used to remove stop words. For example, words "yang", "dapat", "tetapi", "dan", "untuk" were removed from sentences.
- Tokenizing is a process to separate words from the text. Indonesian text separating words by using space characters.
- Duplicate Word Removal is a process to remove duplicate words. For example, word "volatile" and "hilang" has two occurrences in reference answers.

TABLE II. PREPROCESSING

	Student Answer	Reference Answer
Case folding	volatile merupakan memory yang datanya dapat ditulis serta dihapus tetapi akan hilang jika tidak ada aliran listrik kondisi mati off dan membutuhkan daya untuk mempertahankan memory	volatile memory adalah memory yang datanya dapat ditulis dan dihapus tetapi hilang saat kehilangan power kondisi off atau mati lampu
Stemming	volatile rupa memory yang data dapat tulis serta hapus tetapi akan hilang jika tidak ada alir listrik kondisi mati off dan butuh daya untuk tahan memory	volatile memory adalah memory yang data dapat tulis dan hapus tetapi hilang saat hilang power kondisi off atau mati lampu
Stopword Removal	volatile rupa memory data tulis hapus akan hilang tidak alir listrik kondisi mati off butuh daya tahan memory	volatile memory memory data tulis hapus hilang hilang power kondisi off mati lampu
Tokenizing	'volatile', 'rupa', 'memory', 'data', 'tulis', 'hapus', 'akan', 'hilang', 'tidak', 'alir', 'listrik',	'volatile', 'memory', 'memory', 'data', 'tulis', 'hapus', 'hilang', 'hilang', 'power', 'kondisi', 'off', 'mati', 'lampu'

	Student Answer	Reference Answer
	'kondisi', 'mati', 'off', 'butuh', 'daya', 'tahan', 'memory'	
Duplicate Removal	'volatile', 'hilang', 'rupa', 'tidak', 'off', 'akan', 'butuh', 'daya', 'memory', 'data', 'kondisi', 'tahan', 'hapus', 'tulis', 'alir', 'listrik', 'mati	'volatile', 'hilang', 'off', 'power', 'memory', 'data', 'kondisi', 'lampu', 'hapus', 'tulis', 'mati'

C. Reference Answer Word Expansion

Reference answer word expansion in this study is a process of adding word synonyms obtained from Indonesian thesaurus to reference answer word list that is extracted from pre-processing step. Fig. 2 shows expansion using the thesaurus. Each word in the reference answer list is checked if synonym(s) are in the thesaurus. If synonym(s) exist, then a list of word synonym(s) is added to the reference answer list. For example, a list of words ["tulis"], a synonym of "tulis" is "catat", then an expanded list of words ["tulis","catat"]. These synonyms of word expansion are used as an alternative word to enhance similarity between reference answers and student answers.

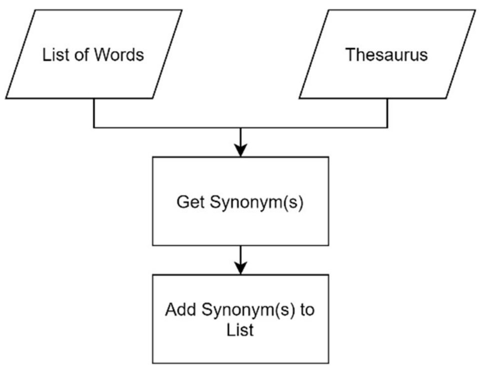

Fig. 2. Word Expansion using Thesaurus

D. Calculate Similarity

String-based similarity methods used in this study to measure similarity between a set of terms in reference answer sentences (a_1) and set of terms in student answer sentences (a_2) are Cosine Coefficient (CC), Jaccard Coefficient (JC), and Dice Coefficient (DC) [9], [10]. Basically, the similarity between these two answers is provided by their intersection words. Note that we only use the thesaurus as alternate word. The number of words counted in this study are gained from the original reference answer (a_1).

The Cosine Coefficient (CC) or Cosine Similarity is a method used to calculate the similarity between two objects using the cosine equation. The similarity between reference answers and student answers was calculated by (2) [15]; however, in this study, we used string-based similarity [10]. As an example from TABLE II. after duplicate removal, we have 17 words students answer, 11 words reference answer, and intersection between these two answers is 9 words. Similarity score for CC is 9/(3.32 * 4.12) = 0.66.

$$score\_sim_{CC}(a_1, a_2) = |a_1 \cap a_2| \,/\, (|a_1|^{1/2} \,.\, |a_2|^{1/2}) \quad (2)$$

Jaccard Coefficient (JC) is a method used to calculate the similarity between two objects by counting the number of the same terms compared by the number of unique terms from two strings [10][16]. JC equation can be seen in (3). For example, as in CC, JC similarity is 9/(11+17-9) = 0.47

$$score\_sim_{JC}(a_1, a_2) = |a_1 \cap a_2| / (|a_1| + |a_2| - |a_1 - a_2|) \quad (3)$$

The dice coefficient is a method for comparing two different text samples. The difference between this method and the Jaccard coefficient is that the dice coefficients did not count the number of unique terms unless counting all terms. The dice coefficient equation can be seen in (4). The dice coefficient is defined as twice the number of the same terms in the two terms being compared, divided by the total number of terms in both texts a_1 and a_2 [9], [10], [17]. For example, as in CC, DC similarity score is 2*9/(11+17) = 0.64

$$score\_sim_{DC}(a_1, a_2) = 2|a_1 \cap a_2| / (|a_1| + |a_2|) \quad (4)$$

E. Keyword Matching

The keyword matching process is carried out to see the match between reference answers and student answers where keywords are obtained by taking unique words from preprocessing stage. The assessment in the keyword matching process is calculated based on the match between student answers against reference answers. This method only consider the intersection of two answers compared to the reference answer (5).

$$Score_{Matching}\,(a_1, a_2) = |a_1 \cap a_2| / |a_1| \quad (5)$$

As an example in TABLE II. after duplicate removal, there are 17 words students answer, and 11 words reference answer. Thus, the intersection between these two answers is 9 words. Keyword matching score is 9/11 = 0.82. In Keyword Matching scoring, the student answers will be worth 0 if no words match the reference answer, and 1 if all the words in the reference answer are in the student's answer

F. Calculate Final Score

The final score is obtained by multiplying similarity score and keyword matching score that ranges [0,1] by 100 (maximum score for each correct answer) to obtain a predicted score. The final score is calculated by averaging predicted scores from similarity and scores from keyword matching. The equation for final score can be seen in (6) or (7).

$$Finalscore_{sim} = 0.5((score_{sim}*100) + (score_{matching}*100)) \quad (6)$$

$$Finalscore_{sim} = 50(score_{sim} + score_{matching}) \quad (7)$$

For example, the final score for similarity using similarity (CC, JC, DC) and keyword matching will be:

$Finalscore_{CC}$ = 50(0.66+0.82) =74

$Finalscore_{JC}$ = 50(0.47+0.82) =64.5

$Finalscore_{DC}$ = 50(0.64+0.82) =73

G. Performance Evaluation

System performance is evaluated by calculating the Mean Absolute Error (MAE) value which can be seen in (8) [17] that is error between the scores obtained by the scoring system and average scores given by three teachers.

$$MAE = \sum |X - Y| \,/\, n \quad (8)$$

Where X is the system's scoring results, Y is the average of students' score answers given by three teachers, and n is the total number of answers evaluated. In addition to MAE, performance evaluation is also carried out by looking for correlations. This correlation is used to find the level of agreement or suitability between student answer scores given by human evaluators and the scores obtained from the automated essay scoring system. The correlation value in this study uses the Pearson Correlation, which can be seen in (9). Correlation is defined as a comparison between covariance and multiplication of standard deviation of student scores given by teachers and scores generated by the system.

$$Corr(X,Y) = Covariance(X,Y)/(stdev(X)*stdev(Y)) \quad (9)$$

Corr is Pearson correlation value between score generated by system and score given by teacher and defined as comparison between covariance and *stdev*. Correlation range of values between 0 to 1. The correlation criteria to determine the system's success is less if the correlation is < 0.4, good if the correlation is between 0.4 and 0.75, and very good if the correlation value is > 0.75 [21].

III. Experimental Result

The experiment design in this study focuses on investigating performance improvement after reference answer word expansion. Mean Absolute Error (MAE) and correlation calculated for each question in categories. We want to achieve smaller MAE and higher correlation.

TABLE III. shows MAE for technology category for Final Score similarity using Cosine Coefficient (CC), Jaccard Coefficient (JC), and Dice Coefficient (DC). As we can see, there are improvement results by expanding reference answers. MAE shows in the table are for score in range [0, 100] the smaller MAE the better. Maximum MAE difference is 5.2931 in technology category is MAE for question 1 using JC similarity from 11.5863 to 6.2932. Overall most significant change is in question 4, also maximum MAE difference for CC and DC similarity. MAE for questions 7 and 8 did not change, indicating that for these 2 questions, reference answer synonym has no intersection with student answer.

TABLE III. Mean Absolute Error (MAE) for Technology Category

Question	Technology			Technology Expanded Answer		
	CC	***JC***	***DC***	***CC***	***JC***	***DC***
1	10.4991	11.5863	10.4247	10.4422	6.2932	9.5391
2	10.5001	14.7365	11.5361	7.0955	11.0582	7.7226
3	11.0833	19.6427	12.6679	9.3890	18.0380	10.9860
4	18.8138	23.9216	18.9433	13.9849	18.8443	13.9084
5	6.5682	7.1544	5.9413	5.6737	5.3671	4.8077
6	11.1905	16.7713	11.3559	11.0613	16.6664	11.2359
7	5.0349	6.1462	4.8559	5.0349	6.1462	4.8559
8	13.7185	15.3055	13.4628	13.7185	15.3055	13.4628
9	12.5452	18.6742	13.5710	10.7519	16.8437	11.5515
10	13.7334	18.6620	14.4257	11.4652	16.2454	11.9690
Average	11.3687	15.2601	11.7185	9.8617	13.0808	10.0039

Table IV shows MAE for all question categories. Word expansion using synonyms increase scoring accuracy in all categories. However, the lifestyle category has the most significant changes when compared before and after using synonyms. The difference for lifestyle category each successive similarity for CC, JC, and DC is 2.3108, 2.3527, and 2.3891. Smallest MAE for all category achieved by CC Final score, followed by DC and JC.

TABLE IV. Mean Absolute Error (MAE) for ALL Categories

Category \ Similarity		CC	JC	DC
Lifestyle	***Base***	11.8903	14.6270	12.3705
	Expanded	9.5794	12.2744	9.9814
Politic	***Base***	13.2602	16.7163	13.3771
	Expanded	11.9540	15.1047	11.9648
Technology	***Base***	11.3687	15.2601	11.7185
	Expanded	9.8617	13.0808	10.0039
Sport	***Base***	12.1473	16.6172	12.8292
	Expanded	10.8691	14.9220	11.4425

TABLE V. shows Pearson correlation between final score using similarity CC, DC, and JC against average score given by three teachers for technology category. The highest correlation achieved in question 7, by a score of more than 0.95, is a very good correlation. The highest improvement for CC final score lies in Question 2 from 0.2725 to 0.4064 or increase by 0.1339. Furthermore, JC and DC final score in Question 5 increases by 0.0968 and 0.0995 consecutively. Using word expansion, the average correlation in the technology category improved from a good correlation to a very good correlation for JC and DC final scores.

TABLE V. Correlation for Technology Category

Question	Technology			Technology Expanded Answer		
	CC	***JC***	***DC***	***CC***	***JC***	***DC***
1	0.6368	0.6982	0.6694	0.6969	0.7594	0.7392
2	0.2725	0.4859	0.4887	0.4064	0.5737	0.5771
3	0.7036	0.7710	0.7892	0.7405	0.7920	0.8121
4	0.5239	0.6364	0.6349	0.6214	0.7040	0.6958
5	0.5688	0.6611	0.6662	0.6803	0.7579	0.7656
6	0.9089	0.9152	0.9189	0.9118	0.9168	0.9214
7	0.9552	0.9509	0.9616	0.9552	0.9509	0.9616
8	0.7568	0.7230	0.7790	0.7568	0.7230	0.7790
9	0.5479	0.6309	0.6534	0.6073	0.6673	0.6949
10	0.6346	0.7019	0.7023	0.5456	0.6621	0.6547
Average	0.6509	0.7174	0.7264	0.6922	0.7507	0.7602

Correlation for all categories can be shown in TABLE VI. Without word expansion, very good correlation type only in Sport category using DC Final score. After expansion, we had 2 categories with a very good correlation in JC similarity and 4 (all) categories with very good correlation. Final score using CC Final score all lies in good category. The highest correlation for all questions category lies in DC Final score, followed by JC and CC both with and without word expansion.

TABLE VI. Correlation for All Category

Category \ Similarity		CC	JC	DC
Lifestyle	***Base***	0.6533	0.7161	0.7212
	Expanded	0.7185	0.7694	0.7727
Politic	***Base***	0.7054	0.7158	0.7458
	Expanded	0.7214	0.7259	0.7543
Technology	***Base***	0.6509	0.7174	0.7264

	Expanded	0.6922	0.7507	0.7602
Sport	***Base***	0.7106	0.7332	0.7503
	Expanded	0.7159	0.7375	0.7550

IV. CONCLUSION

The Word expansion approach can work well to increase similarity in Indonesian short essay auto scoring, as indicated by increasing correlation and decreasing Mean Absolute Error (MAE) values. The highest correlation was achieved by the Dice Coefficient similarity method by very good correlation, and Smallest MAE was achieved by Cosine Coefficient similarity method.

However, MAE in this study is still high enough for a functional essay auto scoring system. MAE is relatively high, more than 9.5794 for a maximum score of 100. study in this field needed to improve performance. Further study is expected to process the dataset with a more varied question category in short answer that has reference answer.

REFERENCES

[1] H. Rababah and A. T. Al-Taani, "An automated scoring approach for Arabic short answers essay questions," in *ICIT 2017 - 8th International Conference on Information Technology, Proceedings*, Oct. 2017, pp. 697–702. doi: 10.1109/ICITECH.2017.8079930.

[2] F. Rahutomo *et al.*, "Open Problems in Indonesian Automatic Essay Scoring System," *International Journal of Engineering & Technology*, vol. 7, no. 4.44, p. 156, Dec. 2018, doi: 10.14419/ijet.v7i4.44.26974.

[3] R. B. Aji, Z. A. Baisal, and Y. Firdaus, "Automatic Essay Grading System Menggunakan Metode Latent Semantic Analysis E-78 E-79," *Seminar Nasional Aplikasi Teknologi Informasi*, vol. 2011, no. Snati, pp. 1–9, 2011.

[4] A. A. P. Ratna, B. Budiardjo, and D. Hartanto, "Simple: Sistim Penilai Esei Otomatis Untuk Menilai Ujian Dalam Bahasa Indonesia," *MAKARA of Technology Series*, vol. 11, no. 1, Oct. 2010, doi: 10.7454/mst.v11i1.435.

[5] A. Agung and P. Ratna, "SIMPLE-O, the Essay Grading System for Indonesian Language Using LSA Method with Multi-Level Keywords," Oct. 2015. Accessed: Apr. 08, 2021. [Online]. Available: www.iafor.org

[6] A. Amalia, D. Gunawan, Y. Fithri, and I. Aulia, "Automated Bahasa Indonesia essay evaluation with latent semantic analysis," in *Journal of Physics: Conference Series*, Jul. 2019, vol. 1235, no. 1, p. 12100. doi: 10.1088/1742-6596/1235/1/012100.

[7] J. Zeniarja, A. Salam, and I. Achsanu, "Sistem Koreksi Jawaban Esai Otomatis (E-Valuation) dengan Vector Space Model pada Computer Based Test (CBT)," *Seri Prosiding Seminar Nasional Dinamika Informatika*, vol. 4, no. 1, Apr. 2020.

[8] T. Ajitiono and Y. Widyani, "Indonesian essay grading module using Natural Language Processing," May 2017. doi: 10.1109/ICODSE.2016.7936117.

[9] U. Hasanah, A. E. Permanasari, S. S. Kusumawardani, and F. S. Pribadi, "A scoring rubric for automatic short answer grading system," *Telkomnika (Telecommunication Computing Electronics and Control)*, vol. 17, no. 2, pp. 763–770, Apr. 2019, doi: 10.12928/TELKOMNIKA.V17I2.11785.

[10] T. wahyuningsih, H. Henderi, and W. Winarno, "Text Mining an Automatic Short Answer Grading (ASAG), Comparison of Three Methods of Cosine Similarity, Jaccard Similarity and Dice's Coefficient," *Journal of Applied Data Sciences*, vol. 2, no. 2, pp. 45–54, May 2021, doi: 10.47738/JADS.V2I2.31.

[11] F. Rahutomo, Y. P. Putra, and M. H. Ali, "Implementasi Manhattan Distance dan Dice Similarity pada Ujian Esai Daring Berbahasa Indonesia," *Seminar Informatika Aplikatif Polinema*, pp. 171–174, 2019.

[12] M. Jamaluddin, N. Yuniarti, A. Rahmani, and J. Hutahaean, "Aplikasi Penilaian Otomatis Ujian Esai Berbahasa Indonesia Menggunakan Algoritma K-Nearest Neighbor (Studi kasus MAN Cimahi)," *Prosiding Industrial Research Workshop and National Seminar*, vol. 10, no. August 2019, pp. 314–324, Aug. 2020, doi: 10.35313/irwns.v10i1.1404.

[13] Q. Liu, H. Huang, J. Lu, Y. Gao, and G. Zhang, "Enhanced word embedding similarity measures using fuzzy rules for query expansion," *IEEE International Conference on Fuzzy Systems*, Aug. 2017, doi: 10.1109/FUZZ-IEEE.2017.8015482.

[14] M. F. Z. Rounaqi, "Query answering system of Shahih Hadith Muttafaqun 'Alaih using Indonesian Thesaurus based on query expansion and naïve bayes classifier," Nov. 2019.

[15] N. A. Rahman, Z. A. Bakar, and T. M. T. Sembok, "Query expansion using thesaurus in improving Malay Hadith retrieval system," *Proceedings 2010 International Symposium on Information Technology - System Development and Application and Knowledge Society, ITSim'10*, vol. 3, pp. 1404–1409, 2010, doi: 10.1109/ITSIM.2010.5561518.

[16] T. Roshinta, F. R.-P. S. (Seminar, and undefined 2016, "Analisis Aspek-Aspek Ujian Esai Daring Berbahasa Indonesia," *proceeding.sentrinov.org*, Accessed: Aug. 21, 2020. [Online]. Available: http://proceeding.sentrinov.org/index.php/sentrinov/article/view/159

[17] M. Nazief, B. A. A. & Adriani, "Confix- stripping: Approach to Stemming Algorithm for Bahasa Indonesia," *Conferences in Research and Practice in Information Technology Series*, vol. 38, no. 4, 2005.

[18] D. Gunawan, C. A. Sembiring, and M. A. Budiman, "The Implementation of Cosine Similarity to Calculate Text Relevance between Two Documents," *Journal of Physics: Conference Series*, vol. 978, no. 1, p. 012120, Mar. 2018, doi: 10.1088/1742-6596/978/1/012120.

[19] W. H.Gomaa and A. A. Fahmy, "A Survey of Text Similarity Approaches," *International Journal of Computer Applications*, vol. 68, no. 13, 2013, doi: 10.5120/11638-7118.

[20] L. R. Dice, "Measures of the Amount of Ecologic Association Between Species," *Ecology*, vol. 26, no. 3, 1945, doi: 10.2307/1932409.

[21] T. F. de C. Marshall and J. L. Fleiss, "Statistical Methods for Rates and Proportions.," *The Statistician*, vol. 25, no. 1, p. 70, 1976, doi: 10.2307/2988144.