Ministry of Education and Culture Accreditation,
No: 30/E/KPT/2019 (Sinta 4)
DOI: https://doi.org/10.31294/p.v24i2

Paradigma,
Vol. 24, No. 2, September 2022, Page. 175-181
P-ISSN 1410-5063, E-ISSN: 2579-3500

Determining the Best Answers for Balinese Language Problems using Latent Semantic Analysis

Made Agus Putra Subali[1], I Ketut Putu Suniantara[2]

[1,2]Institut Teknologi dan Bisnis STIKOM Bali
Denpasar, Indonesia

e-mail: [1*]madeagusputrasubali@gmail.com

Abstrak - Pada Bahasa Bali, soal uraian atau esai dibentuk dalam format interogatif menggunakan kata tanya seperti *akuda*, *apa*, *dija*, *kenken*, *kuda*, dan *nyen*. Proses penilaian pada soal uraian cenderung lebih sulit dan kompleks dibandingkan dengan soal pilihan ganda, hal ini dikarenakan soal uraian diuraikan dalam bentuk kalimat. Adapun solusi dalam memudahkan proses penilaian pada soal uraian dapat dilakukan dengan menggunakan *automated essay scoring*. Berdasarkan hasil dari penelitian terdahulu, metode *Latent Semantic Analysis* (LSA) memberikan tingkat akurasi yang lebih baik, dikarenakan metode LSA menggunakan metode *Singular Value Decomposition* (SVD) untuk memperoleh pola hubungan baru antara *term* dengan *term* referensi. Data yang digunakan dalam penelitian ini adalah lima soal beserta kunci jawabannya dan terdapat lima kandidat jawaban di setiap soal dalam Bahasa Bali. Berdasarkan hasil pengujian yang telah dilakukan metode yang digunakan memperoleh rerata akurasi pada seluruh soal sebesar 70.26%, hal ini menandakan bahwa metode LSA dapat digunakan dengan baik dalam proses penilaian soal uraian atau *automated essay scoring*.

Kata Kunci: *automated essay scoring*, *latent semantic analysis*, bahasa bali.

Abstract - *In Balinese, descriptions or essays are formed in an interrogative format using question words such as "akuda", "apa", "dija", "kenken", "kuda", dan "nyen". The assessment process on description questions tends to be more difficult and complex than multiple choice questions, this is because the description questions are described in sentence form. The solution to facilitate the assessment process on description questions can be done using automated essay scoring. Based on the results of previous studies, the Latent Semantic Analysis (LSA) method provides a better level of accuracy, because the LSA method uses the Singular Value Decomposition (SVD) method to obtain a new pattern of relationships between terms and reference terms. The data used in this study are five questions and their answer keys and there are five candidate answers for each question in Balinese. Based on the tests that have been carried out, the method used obtained an overall average accuracy of 70.26%, this shows that the LSA method can be used well in the assessment process or automatic essay assessment.*

Keywords*: automated essay scoring, latent semantic analysis, balinese language.*

INTRODUCTION

Essay questions are used to measure the level of understanding of "someone" towards something (Contreras et al., 2018). In Balinese, essay questions are formed in an interrogative format using question words such as *akuda*, *apa*, *dija*, *kenken*, *kuda*, and *nyen* (Granoka et al., 1996). Description or essay questions are described in the form of sentences, this makes the assessment process more difficult and complex compared to multiple choice questions (Chen et al., 2014). Based on previous research, the use of the automated essay scoring method can facilitate the assessment process on description questions and under certain conditions can obtain good accuracy (McNamara et al., 2015).

Previous research related to automated essay scoring has been done by Fauzi, et al. in the e-learning system using the cosine similarity method and the *n*-gram method, the accuracy obtained is 67% (Fauzi et al., 2017). Citawan, et al. on the e-learning system using the latent semantic analysis, cosine similarity, and *n*-gram methods, the accuracy obtained is 78.65% (Citawan et al., 2017).

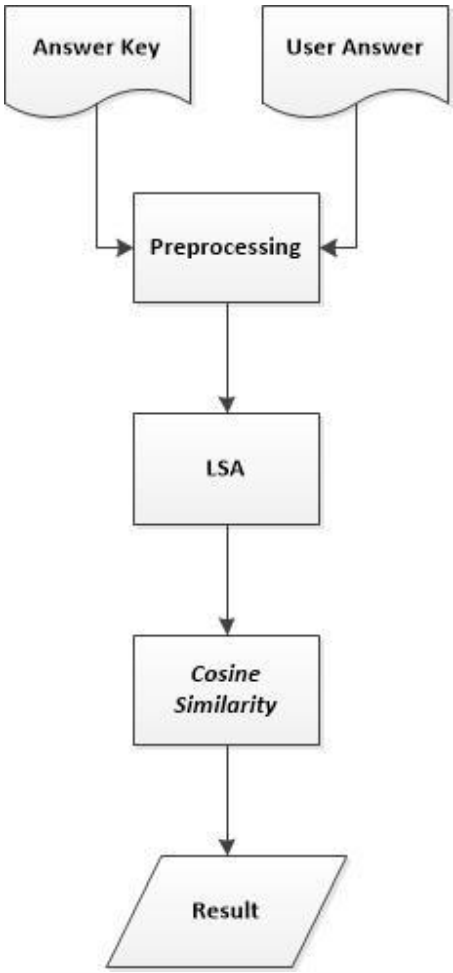
Based on the results of previous studies, the LSA method provides a better level of accuracy, because the LSA method uses the Singular Value Decomposition (SVD) method to obtain a new pattern of relationships between terms and reference terms

Article Information	Received: September 2, 2022	Revised: September 19, 2022	Accepted: September 20, 2022

(Citawan et al., 2017). Considering the process of assessing description questions is more complex than multiple choice questions, we propose a system that helps the process of assessing description questions as well as sorting the answers that are most relevant to the answer key using the LSA method.

RESEARCH METHODOLOGY

Figure 1 is the proposed method.



Source: (Subali & Suniantara, 2022)

Figure 1. Question Answering System Method

1. Answer Key and User Answer

At this stage, two inputs will be given, namely the answer key and the user answer. In the answer key and user answer, the process of changing each character to the lowercase form and removing punctuation marks is carried out.

2. Preprocess

In the pre-processing stage, tokenize, stopwords removal, and stemming processes are carried out. The determination of stopwords in Balinese has been studied by Putra, et al. which includes the words *anggen*, *sane*, *ring*, *miwah*, *puniki*, and *olih* (Putra et al., 2016), and for the Balinese stemming process, we use the stemmer method that has been done by Subali, et al., where the stemmer method uses the rule based method and *n*-gram string similarity (Subali & Fatichah, 2019).

3. LSA

LSA is a method for analyzing the semantic structure of the text by utilizing the statistical computing (Citawan et al., 2017). The following are each step in the LSA method:

a. Form a matrix A, where row i of the matrix contains unique words in each document and column j contains document labels, while the cell contains the frequency of occurrence of words i, j.

b. Applying Singular Value Decomposition (SVD) on matrix A, where the matrix is decomposed into three forms, U, S, and V matrix.

$$A = U \times S \times V^T \quad (1)$$

Information:
$A \in R^{m,n}$
U: matrix *orthogonal*, $U \in R^{m,\min(m,n)}$
S: matrix *diagonal*, $S \in R^{\min(m,n),\min(m,n)}$
V: matrix *orthogonal*, $V \in R^{n,\min(m,n)}$

c. Reduces the matrix by storing all the rows in the first k columns U and V and the first k rows and k columns S.

$$A_k = U_k \times S_k \times V_k^T \quad (2)$$

Information:
k is the number of matrix reduction parameters.

d. To determine the similarity of each text, the matrix obtained by the LSA method will be measured using the cosine similarity method.

4. Cosine Similarity

The cosine similarity method is used to measure the level of similarity between the keywords obtained and the document (Fauzi et al., 2014; Subali & Wijaya, 2021) in equation (3) is a way of measuring the level of similarity using the cosine similarity method.

$$similarity(d_j, q) = \frac{d_j \cdot q}{|d_j| \cdot |q|} = \frac{\sum_{i=1}^{n}(W_{i,j} \cdot W_{i,q})}{\sqrt{\sum_{i=1}^{n} W_{i,j}^2 \cdot \sum_{i=1}^{n} W_{i,q}^2}} \quad (3)$$

Information:
$W_{i,j}$ is the weight of word i in document j.
$W_{i,q}$ is the weight of the word i in the question q.
Term weight is calculated using a bag of words.

RESULTS AND DISCUSSION

1. Data

The research data used were five questions and their answer keys in Balinese. These five questions are topics related to basic computer science. In Table 1 are the five questions used.

Meanwhile, data related to the list of answers will be collected using a questionnaire method, where each question contains five candidate answers. At the time of data collection will also involve five respondents who work as active students.

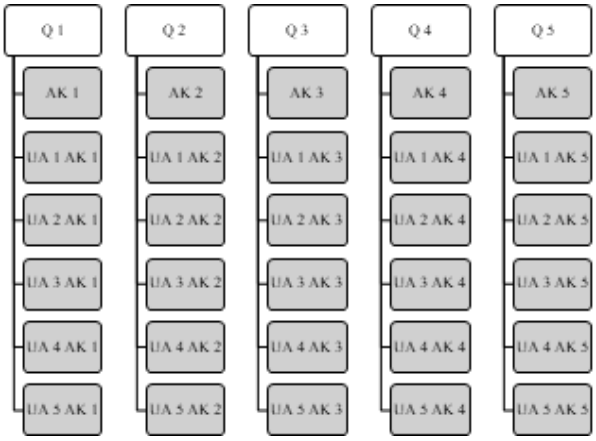
Table 1. Data

No.	Question	Answer Key
1	what is the internet?	the internet is a global computer network infrastructure.
	napi sane kabaos internet?	*sane kabaos internet piranti jaringan komputer global.*
2	what is a web application?	the web or world wide web is one of the services on the internet that serves to provide information via a web browser using the http protocol.
	napi sane kabaos aplikasi web?	*sane kabaos web utawi world wide web pinaka sinalih tunggil layanan sane wenten ring internet sane madue kawigunan ngamolihang informasi majalaran antuk web browser sane ngangge protocol http.*
3	what is the main function of a web browser?	access information on the web.
	napi kawigunan utama web browser?	*mengakses informasi sane wenten ring web.*
4	explain how web applications work?	the user gives a request or request by entering the site address or url through a web browser, then the web browser sends the request to a web server or called an http request, the web server then processes the request to produce a response to be given back to the web browser or called an http response.
	indayang telatarang sapunapi aplikasi web kamargiang?	*pengguna ngicen permintaan utawi request antuk ngetik alamat situs utwi url ring web browser, selanturnyane web browser ngirim permintaan inucap ke web server utawi sane kabaos http request, selanturnyaane web browser memproses request inucap ngantos ngamolihang respon sane jagi kirim malih ke web browser utawi sane kabaos antuk http response.*
5	what is the function of the database?	the database serves to store data in a structured manner in the form of tables.
	napi kawigunan database?	*kawigunan database anggen nyimpen data mangda anut utawi kabaos terstruktur ring bentuk tabel.*

Source: (Subali & Suniantara, 2022)

2. Answer Key and User Answer

The data answer key is the answer key for each question, while the user answer is the answer to the five respondents for each question. Figure 2 shows the answer key (AK) and user answer (UA) initialization model for each question.



Source: (Subali & Suniantara, 2022)

Figure 2. Answer Key and User Answer

If you look at Figure 2, each question number will compare each answer key with all of the respondents' answers to the question.

So we get:

q: $answer\ key_i$

d: $user\ answer_{i,j}$

Information:

i question number.

j respondent number.

3. Preprocess

In the preprocessing stage, the answer key and user answer data are carried out in several stages, starting from tokenize, stop word removal, and finally stemming. In Table 2 is the number of features generated at the data preprocessing stage.

Table 2. Number of Features in All Questions

Question	Number of Features
1	33
2	47
3	27
4	89
5	88

Source: (Subali & Suniantara, 2022)

4. LSA

In the early stages of the LSA method, it is done by forming a term matrix from the answer key and user answer data. Where the term matrix is formed with conditions, where the row contains the features, the column contains the question number, while the cell contains the number of words or features that appear in the question number.

In Table 3, the term matrix in the answer key and user answer in question number one is shown, while the term matrix for other question numbers can be seen at the link https://intip.in/5SoalBesertaRespondenFitur.

Table 3. Term Matrix Question Number One

Features	AK	UA1	UA2	UA3	UA4	UA5
miwah	0	0	0	1	0	0
majalaran	0	1	0	0	0	0
komunikasi	0	0	0	1	0	0
siosan	0	0	1	1	0	0
penghubung	0	0	0	0	0	1
inggih	0	1	1	2	1	1
antuk	0	1	0	0	0	0
terhubung	0	0	0	2	0	1
gumine	0	0	1	0	1	0
global	1	0	0	0	0	0
menghubungkan	0	1	1	0	0	0
media	0	0	0	0	0	1
utawi	0	1	0	1	0	0
satelit	0	1	0	0	0	0
kabaos	1	0	0	0	0	0
ring	0	0	2	0	1	0
perangkat	0	0	2	0	0	0
sistem	0	0	0	0	0	1
punika	0	1	1	2	1	1
makasami	0	0	0	0	1	0
nganggen	0	0	0	0	0	1
antar	0	1	0	0	0	0
piranti	1	0	1	0	0	0
internet	1	1	1	2	1	1
telepon	0	1	0	0	0	0
terkoneksi	0	0	0	0	1	0
jaringan	1	1	1	3	1	1
majeng	0	0	1	0	0	0
anggen	0	0	1	1	0	0
sane	0	1	1	2	1	1
computer	1	1	0	2	1	0
sareng	0	0	0	1	1	0
informasi	0	0	0	1	0	0

Source: (Subali & Suniantara, 2022)

Information:
AK is answer key.
UA is user answer.

There were 33 features obtained from the answer key and the five user answers to question number one.

After the term matrix is obtained, then the matrix decomposition process is carried out using the SVD method which produces three different matrices, namely the U, S, and V^T matrices using equation (1). From the three matrices, the matrix reduction process is then carried out by storing all rows in the first k columns U and V and the first k rows and k columns S using equation (2), where the value of $k = 2$.

In Table 4, the decomposition matrix of U_k in question number one is obtained.

Table 4. Decomposition Matrix U_k in Question Number One

	0	1
0	-0,04548441	-0,24512673
1	-0,13198261	-0,53191654
2	-0,04085392	0,02419411
3	-0,35480696	-0,00560556
4	-0,08633834	-0,22093262
5	-0,04085392	0,02419411
6	-0,04085392	0,02419411
7	-0,14141712	-0,13509740
8	-0,04101378	-0,04166308
9	-0,13678663	0,13422344
10	-0,03558942	0,03693148
11	-0,13694649	0,06836625
12	-0,09593271	0,11002933
13	-0,03558942	0,03693148
14	-0,35480696	-0,00560556
15	0,00000000	0,00000000
16	-0,04101378	-0,04166308
17	-0,45073966	0,10442377
18	-0,04085392	0,02419411
19	-0,09096883	-0,49025345
20	-0,03558942	0,03693148
21	-0,09593271	0,11002933
22	-0,04085392	0,02419411
23	0,00000000	0,00000000
24	-0,35480696	-0,00560556
25	-0,04548441	-0,24512673
26	-0,09593271	0,11002933
27	-0,35480696	-0,00560556
28	-0,27373312	0,20258968
29	-0,03558942	0,03693148
30	-0,14141712	-0,13509740
31	-0,08649819	-0,28678981
32	-0,22745484	0,25699014

Source: (Subali & Suniantara, 2022)

In Table 4 it can be seen that the U_k matrix only takes 2 columns from the U decomposition matrix.
In Table 5 it is a decomposition matrix of S_k, while in Table 6 it is a decomposition matrix of V_k^T in question number one.

Table 5. S_k Decomposition Matrix in Question Number One

	0	1
0	7,933087	0
1	0	3,630280

Source: (Subali & Suniantara, 2022)

Table 6. V_k^T Decomposition Matrix in Question Number One

	0	1	2	3	4
0	-0,3241	-0,36083	-0,76104	-0,32537	-0,28233
1	0,087831	-0,88988	0,399437	-0,15125	0,134072

Source: (Subali & Suniantara, 2022)

In Table 6, values are obtained for each vector of respondents' answers to question number one, where:
$UA1$: (-0.32410, 0.087831)
$UA2$: (-0.36083, -0.88988)
$UA3$: (-0.76104, 0.399437)
$UA4$: (-0.32537, -0.15125)
$UA5$: (-0.28233, 0.134072)

Information:
UA is user answer.

5. Cosine Similarity
Before the similarity measurement process is carried out, the vector value for the answer key must first be calculated using equation (4), as follows:

$$AK = AK^T \cdot U_k \cdot S_k^{-1} \quad (4)$$

Information:
AK is answer key.
AK^T is answer key transpose.
U_k is the decomposition matrix U.
S_k^{-1} is the decomposition matrix S power -1.

In Table 7 is the vector value of the answer key in question number one.

Table 7. Vector Value of the Answer Key to Question Number One

	0
0	-0,141781402
1	0,015503256

Source: (Subali & Suniantara, 2022)

So that the vector value in the answer key to question number one is obtained, namely:
AK: (-0.141781402, 0.015503256)

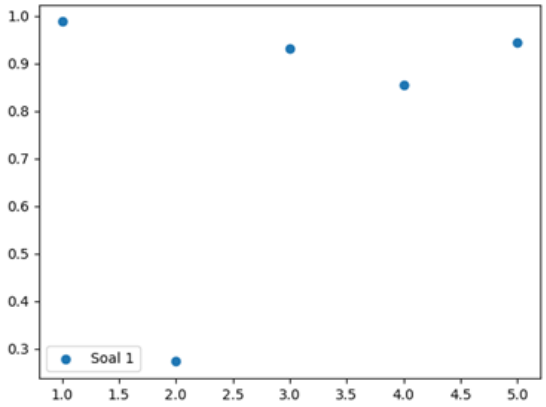
Information:
AK is answer key.

The last step is to calculate the level of similarity between the answer keys and each respondent's answer using the cosine similarity in equation (3). The following is the process of measuring the level of similarity in question number one for respondent number one, $similarity(d_1, q)$ is 0.9879 or 98.79%.

$$= \frac{(-0.3241)(-0,141781402) + (0.087831)(0,015503256)}{\sqrt{(-0.3241)^2 + (0.087831)^2}\sqrt{(-0,141781402)^2 + (0,015503256)^2}}$$

$$= 0.98789814$$

Figure 3 is the result of calculating the level of similarity of all respondents in question number one.



Source: (Subali & Suniantara, 2022)
Figure 3. Scatter Plot Calculation of the Similarity Level of All User Answers and Answer Key in Question One

6. Results
Details of the results of the level of similarity in all question numbers can be seen in Table 8.

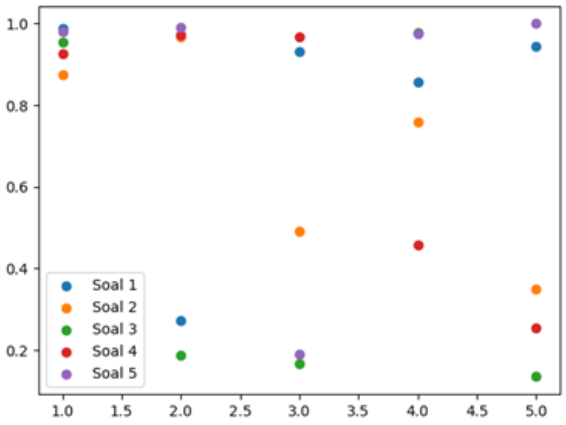
Table 8. The Results Similarity Level of All Questions

Question	User Answer	Similarity Score
1	Respondent 1	**0.98789814**
	Respondent 2	0.27280914
	Respondent 3	0.93071983
	Respondent 4	0.85561745
	Respondent 5	0.94459825
	Average	**0,79832856**
2	Respondent 1	0.87514644
	Respondent 2	**0.96623994**
	Respondent 3	0.49018146
	Respondent 4	0.75921970
	Respondent 5	0.35052616
	Average	**0,68826274**
3	Respondent 1	0.95335390
	Respondent 2	0.18803289
	Respondent 3	0.16539411
	Respondent 4	**0.97652961**
	Respondent 5	0.13459548
	Average	**0,48358120**
4	Respondent 1	0.92682068
	Respondent 2	**0.97241178**
	Respondent 3	0.96792410
	Respondent 4	0.45799945
	Respondent 5	0.25441755
	Average	**0,71591471**

5	Respondent 1	0.97972970
	Respondent 2	0.99116090
	Respondent 3	0.18960860
	Respondent 4	0.97495892
	Respondent 5	**0.99995648**
	Average	**0,82708292**

Source: (Subali & Suniantara, 2022)

Figure 4 shows all the results of the comparison between the answer keys and the user's answers for each question.



Source: (Subali & Suniantara, 2022)
Figure 4. Scatter Plot Calculation of the Similarity of All Questions

Based on the results of the level of similarity obtained and a manual examination of the answers of each respondent and the answer key showed very good accuracy (close to the maximum value of one) for the five questions.
It can be seen from the average value obtained in each question, that the majority obtained an average value of > 0.5, or the average accuracy on all questions was 0.70263403 or 70.26%. This proves that the LSA and cosine similarity methods can be applied well in the process of automatically scoring essay questions.
If you look at Figure 4, it is only in question number three that the majority get an average value of < 0.3, this is due to the lack of similarity in word structure in respondent's number 2, 3, and 5 when compared to the answer key, on the other hand, respondents with numbers 1 and 4 get accuracy. which is very good.
In addition, in another case in question number 5 respondent 3 obtained a very small accuracy of 0.18960860 compared to other respondents on the same question, this is because the number of dimensions or features in respondent 3 is very large compared to the number of features in the answer key, this is an advantage of the LSA method because the LSA method in addition to paying attention to the structure of the occurrence of similar words, the LSA method also pays attention to the number of data features being compared.

CONCLUSION

The application of the LSA and cosine similarity methods to the automated essay scoring that has been carried out has obtained a good average accuracy of 70.26% of tests that have been carried out. Based on the test results, the LSA method is also able to overcome the difference in the number of features in the words being compared, this is because the LSA method does not only focus on paying attention to the structure of the similarity of words but also pays attention to the number of features of the words being compared.
The LSA and cosine similarity methods have a weakness when the compared word conditions have the same word similarity structure but have different word orders, then the LSA and cosine similarity methods will provide a high level of similarity value even though the word order has differences or does not have meaning. In future research, the LSA and cosine similarity methods will use the *n*-gram method in the formation of each feature to be able to focus on paying attention to the word order when two words are compared.

REFERENCES

Chen, H., Xu, J., & He, B. (2014). Automated Essay Scoring by Capturing Relative Writing Quality. *The Computer Journal*, *57*(9), 1318–1330.

Citawan, R. S., Mawardi, V. C., & Mulyawan, B. (2017). Automatic Essay Scoring in E-learning System Using LSA Method with N-Gram Feature for Bahasa Indonesia. *International Conference on Electrical Systems, Technology and Information (ICESTI)*.

Contreras, J. O., Hilles, S., & Abubakar, Z. B. (2018). Automated Essay Scoring with Ontology based on Text Mining and NLTK Tools. *International Conference on Smart Computing and Electronic Enterprise (ICSCEE)*.

Fauzi, M. A., Arifin, A. Z., & Yuniarti, A. (2014). Term Weighting Berbasis Indeks Buku dan Kelas untuk Perangkingan Dokumen Berbahasa Arab. *Lontar Komputer*, *5*(2), 435–442.

Fauzi, M. A., Utomo, D. C., Setiawan, B. D., & Pramukantoro, E. S. (2017). Automatic Essay Scoring System Using N-Gram and Cosine Similarity for Gamification Based E-Learning. *International Conference on Advances in Image Processing (ICAIP)*.

Granoka, I. W. O., Naryana, I. B. U., Jendera, I. W., Bawa, I. W., Medera, I. N., Putrayasa, I. G. N., Anom, I. G. K., Tama, I. W., Denes, I. M., Purwa, I. M., Sukayana, I. N., & Indra, I. B. K. M. (1996). *Tata Bahasa Baku Bahasa Bali*.

Balai Penelitian Bahasa Pusat Pembinaan dan Pengembangan Bahasa Departemen Pendidikan dan Kebudayaan.

McNamara, D. S., Crossley, S. A., Roscoe, R. D., Allen, L. K., & Dai, J. (2015). A hierarchical classification approach to automated essay scoring. *Assessing Writing*, *23*, 35–59.

Putra, I. B. G. W., Sudarma, M., & Kumara, I. N. S. (2016). Klasifikasi Teks Bahasa Bali dengan Metode Supervised Learning Naive Bayes Classifier. *Teknologi Elektro*, *15*(2), 81–86.

Subali, M. A. P., & Fatichah, C. (2019). Kombinasi Metode Rule-Based dan N-Gram Stemming untuk Mengenali Stemmer Bahasa Bali. *Jurnal Teknologi Informasi Dan Ilmu Komputer (JTIIK)*, *6*(2).

Subali, M. A. P., & Suniantara, I. K. P. (2022). *Determining the Best Answers for Balinese Language Problems using Latent Semantic Analysis*.

Subali, M. A. P., & Wijaya, P. (2021). Sistem Question Answering untuk Bahasa Bali menggunakan Metode Rule-Based dan String Similarity. *Techno.COM*, *20*(2), 300–308.