JSE Journal of Science and Engineering, Vol. 2, No. 2 (page: 95-100)
e-ISSN: 2723-3944 January 2024
DOI: https://doi.org/10.30650/jse.v1i2.3841

Indonesian Automated Essay Scoring with Bag of Word and Support Vector Regression

Naufal Azmi Verdikha[1 *], Junianda Haris Dwiagam[2], Rofilde Hasudungan[3]
[1,2,3] Universitas Muhammadiyah Kalimantan Timur, Samarinda, Indonesia
* Corresponding Email: nav651@umkt.ac.id

Abstract – Essay is one of the test questions to measure students' understanding of learning. Respondents can organize the answers to each question in their own language style, so it takes time to make corrections. It takes a system that can assess essay answers automatically quickly and accurately. Auto Essay Scoring (AES) is a tool that can assign grades or scores to answers in the form of essays automatically. In giving grades automatically, AES requires machine learning with training data that contains answer data that has been given a value by the assessor. In this study, AES was used to assess the Indonesian language midterm exams using the Bag of Word extraction feature and using Support Vector Regression. The Root Mean Square Error value obtained when evaluating AES is 1.99.

Keywords: Automated Essay Scoring; Bag of Word; Support Vector Regression

Submitted: 22 November 2023 - Revised: 29 November 2023 - Accepted: 12 January 2024

1. Introduction

Essay is a test in the form of structured questions, the answerer or test will arrange and organize independently the answers to each question with their own style of language. Answers can be limited short answers that have been specified such as name, place, or time, and there are also free answers that require a point of view or opinion from the respondent's side where this answer allows for a variety of different answers. The variety of answers to the questions given makes it difficult when making corrections, and allows for inaccuracies when making assessments because of the large number of essays that must be corrected, thus tiring the proofreaders. In terms of performance, lecturers need extra time and energy to check student answers one by one, with an estimated time of 5 minutes for 1 answer sheet. While the time you have for corrections is not much. The process of proofreading is also often disturbed by other factors outside the measurement intent such as beauty, neatness of writing and also subjectivity. A system is needed that can help evaluate the answers to essay questions automatically so that assessments can be completed quickly and accurately[1] .

Auto Essay Scoring or abbreviated as AES, a tool whose job is to assign a grade or score to a question or quiz in the form of an essay, with the aim of reducing the work of human involvement in terms of assessment. This system is an automatic assessment system and natural language processing or known as Natural Language Processing[2]. AES has been implemented a lot to complete in terms of scoring or automatic scoring on essay questions. In 2015 Yustiana implemented AES using the Latent Semantic Analysis (LSA) and Euclidean Distance methods, where LSA evaluates the similarity of words and Euclidean Distance measures the similarity between answer keys and answers given by students. In research, the results of the tests conducted showed that this system was able to carry out an automatic assessment process for Indonesian language essay answers [3].

There are four types of AES, which are widely used by testing companies, universities and public schools, including Project Essay Grader (PEG) which was the first system created to grade essays, Intelligent Essay Assessor (IEA) which is an essay grading system which uses the Latent Semantic Analysis method, E-rater which is used by the Educational Testing Service to grade essays on the Graduate Management Admissions Test, and IntelliMetric, AES developed by Vantage Learning and used by the College Board.[4]

In 2021 Thamrin conducted research on the automation of assessments on an Indonesian-language student essay answer in an online exam system at Muhammadiyah University, East Kalimantan [5]. Thamrin uses Tf-Idf as a text data extraction feature, then the text data extraction results will then be processed by Text Classification. Prior to the Text Classification, Thamrin conducted a Split Data Text, namely dividing the Training Data and also the Test Data. Thamrin uses KNN and

SVM Classifier in Text Classification. RMSE (Root Mean Square Error), an evaluator whose function is to calculate the number of errors or errors that occur in the algorithm, the difference from the estimated true value and find the average number of squares of errors. From the extraction and classification results that have been carried out by Thamrin, then enter the RMSE evaluation stage to calculate the error values obtained from the extraction and evaluation results that have been carried out. The research conducted by Verdikha (2021) made improvements to the research conducted by Thamrin (2021). Verdikha conducted an experiment to determine the proper performance of the learning model by making comparisons using the SMOTE oversampling method and some of the best regression methods using Support Vector Regression (SVR), Logistic Regression (LR), and Multi-Layer Perceptron Regressor (MLP-R). By using the TF-IDF extraction feature as a text data extraction feature, and RMSE as an evaluation value, the evaluation value is used to compare the results of previous studies [6].

Based on the research researched by Thamrin (2021), it shows that the SVM classification method shows a better RMSE value compared to the KNN and KNN classification methods using LSA. Verdikha's research (2021) compares and gains improvement to Thamrin's research (2021). In his research, Verdikha shows that the Regression SVR method has a better RMSE than the SVM Thamrin method. The RMSE value for each method is 2.730 for RMSE Thamrin using SVM classification and 2.166 for RMSE Verdikha using SVR regression, which shows a lower RMSE value. The SVR regression method has been shown to reduce the RMSE value compared to the SVM classification method.

Research by Thamrin (2021) & Verdikha (2021) uses the Tf-Idf feature as a text data extraction feature. There are many features that can be used to extract data, especially in text, one of which is Bag of Word (BoW). The BoW feature has been used to solve text data extraction cases such as in research conducted by Prasetyo (2017). Prasetyo classifies reviews on a mobile app store by utilizing the Github Issue Tracker. The BoW data extraction process produces unigram and bigram Bag of Words output, which will then be classified and tested and evaluated. Unigram bag of words produces better accuracy than bigram bag of words (Prasetyo, 2017). Based on the previous studies cited above, the authors can draw several conclusions that the use of the BoW feature helps in extracting data and calculating text frequency weights. From the research cited above, BoW can be used to extract text, such as an application review conducted by Prasetyo (2017). In Verdikha's research (2021) it shows that the SVR regression method can be used in carrying out regressions with good evaluations.

Based on the previous research above, this study conducted further research using the Bag of Words (BoW) extraction feature in implementing an AES on the data used in Thamrin's (2021) and Verdikha's (2021) studies using the SVR method as Regression and RMSE as learning module evaluator.

2. Related Works

The implementation of AES was applied to Sharma & Jayagopi's research (2018). Researchers automate grading of handwritten essays by incorporating OHR (Optical Handwriting Recognition and AES) systems. AES are trained to use the GloVe word vector feature. Scores from all essays are collected from AES and OHRT, as well as manually transcribed essays. The OHR system can be used to applications such as AES [7].

The use of AES by utilizing Deep Learning integration, in Lu & Cutumisu's research (2021). Researchers integrate Deep Learning into an Automatic Feedback Generation System on Automated Essay Scoring. Researchers implement, compare, and contrast three AES algorithms (CNN, LSTM, and BiLSTM) with word-embedding and deep learning models. Researchers proved the accuracy of the assessment using the AES algorithm outperforms the latest, most recent models, and the CGMH method produces semantically related feedback sentences [8].

Subsequent research uses the Backpropagation Neural Network method with Lexicon Based Features combined with Bag of Words. The results of the comparison of the Backpropagation Neural Network method based on Lexicon Based Features and Bag of Words are not better than the Random Forest Decision Tree using n-gram features in previous research.[9].

3. Research Methods

This study looked for the RMSE evaluation value in AES in Indonesian using the Bag of Words data extraction method and Support Vector Regression as for regression job for predicting the score of data. The framework for the research stages can be seen in Figure 1.

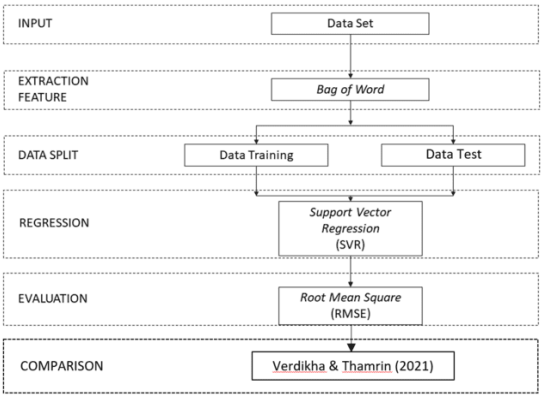

Figure 1. Research Method

e-ISSN: 2723-3944

Further explanation regarding the research stages in Figure 1 is as follows:

1. Input, perform dataset input
2. Extraction Feature, data extraction uses the Bag of Word feature which is implemented using Scikit-Learn.
3. Data Split, split dataset to data training and Data Test
4. Regression, using the Support Vector Regression (SVR) method which is implemented using Scikit-Learn.
5. Conduct learning module analysis to determine the use of the SVR method with polynomial kernel parameters using Root Mean Square Error (RMSE).

This study uses data obtained from research conducted by Thamrin & Verdikha (2021). This data is taken from the results of Indonesian language essay answers conducted on a campus in semester 2 students in 2020 where the data consists of 1648 rows and 3 columns, each column containing grades, class, and student answers. Distribution data is shown in Figure 2. One of the student answer data can be seen in Figure 3.

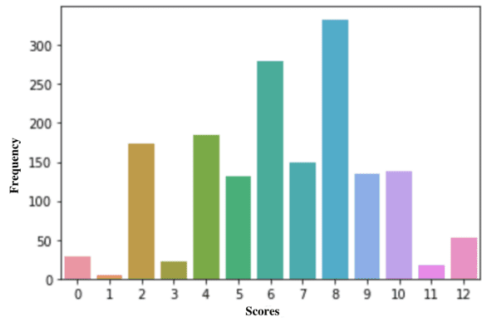

Figure 2. Distribution Data

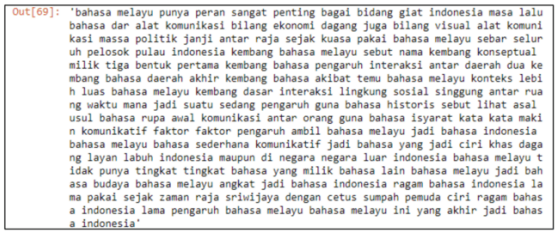

Figure 3. One of The Student Answer Data

Raw data or often referred to as raw data, pre-processing steps are needed before implementing Machine Learning because the algorithm learns from data, and learning outcomes depend heavily on the right data. Before carrying out learning it is necessary to remove noise or distractions, normalize data that is not normal, and also pre-processing is needed to reduce the size of data that is too large, because data that is too large can become a distraction because there is something unnecessary/relevant in doing Machine Learning.

This dataset was obtained from previous research. The answer texts in this dataset have gone through several pre-processing stages in previous research. The pre-processing stages that have been carried out as follows:

- Data Cleaning: Correct or delete data that is not needed in the data frame which consists of the value answer dataset, class, and answer. Data Cleaning includes dropping Columns on the Data Frame, changing the Index on the dataframe, and tidying up the Fields on the data.
- Case Folding: Change all letters in the answer text to lowercase.
- White Space Removal: Eliminate or remove whitespace.
- Spell Correction: Fix or correct the spelling.
- Stopword Removal: Filtering important words from Stopwords, namely words that have low information from a text. Examples ("yang", "dan", "di", "dari", etc).
- Stemming: Change affixed words to basic forms.

All documents can be represented simply using Bag of words (BoW). BoW is a model that represents objects globally, for example text sentences or documents as bag (multiset) words regardless of grammar and even word order to maintain their diversity. In other words, BoW is a collection of unique words in a document. Bag of Word processes each document that has been input by calculating the number of times each word appears. BoW ignores the word order in each document, the syntactic structure of documents and sentences [10].

The establishment of the Bag of Words extraction feature in determining document text can be exemplified in a simple manner as follows:

Text : "*Samara suka mengambil foto pemandangan, Ayu juga suka melukis pemandangan"*

From the text of the statement above it can be arranged into a BoW by using unique words that appear represented only once, thus forming a different unique word order, and then counting the number of times the unique words appear

Table 1
Example Formation of Bag of Words.

No	Words	Frequency
1	Samara	1
2	suka	2
3	mengambil	1
4	foto	1
5	pemandangan	2
6	Ayu	1
7	juga	1
8	melukis	1

This study use Scikit Library (sklearn) for feature extraction using Bag of Word. In Sklearn, the Bag of Word extraction feature can be implemented using the CountVectorizer module. In sklearn, Bag of Words is a CountVectorizer, Count vectorizer creates a matrix with documents and a number of tokens (bag of terms/tokens).

e-ISSN: 2723-3944

CountVectorizer implements tokenization and count occurrences in one class.

After extracting the text data, the dataset is divided randomly into training data and test data, with a data comparison ratio of 8:2. This comparison ratio is taken from previous research conducted by Thamrin (2021) & Verdikha (2021) which used an 8:2 ratio to perform split data. Ratio 8 for training data as the fit model and test data, then ratio 2 for test data as evaluator data from the results of the fit model.

In the Split Data stage, training data and test data are determined randomly using the random state feature. This random state is a parameter whose job is to initialize the generator randomly, which decides the selected data to be used as training data or test data. In this study, the random state used was the random 42 pattern, which was used in previous studies by Thamrin (2021) and Verdikha (2021) which also used random state 42 . This random state pattern produces random patterns of training data and test data that are the same as those obtained in previous studies.

The regression and evaluation used are SVR and RMSE as evaluator. SVR is a development algorithm from the theory of applying the Machine Learning Support Vector Machine (SVM) in the Regression case which produces real and continuous output numbers. [11]. SVR has the ability to overcome overfitting problems, so it can get a function with a small error rate and produce good predictions. RMSE is a parameter that is used to evaluate the value of the results of measurements against the actual value or the value that is considered correct. This RMSE is an evaluation stage in modeling data by minimizing the error rate or the difference between the predicted value and the actual value and evaluating the performance of the algorithm / used. The smaller the RMSE value, the closer the data clustering is to true. RMSE calculates the error (or difference) of the forecast to the actual value and averages the sum of the squares of the errors.

4. Results and Discussion

The results of importing data from student answers that have been collected in a file in the form of CSV (Comma Separated Values), the results of the data that have been imported are shown as follows:

Out[48]:

	nilai	kelas	jawaban
0	8	A	4 faktor yg sebab sebab bahasa melayu angkat j...
1	8	A	bahasa melayu angkat jadi bahasa satu indonesi...
2	8	A	empat faktor sebab bahasa melayu angkat jadi b...
3	8	A	alas bahasa melayu pilih jadi bahasa indonesia...
4	6	A	bahasa indonesia tumbuh kembang bahasa melayu ...
...	...	...	...
1643	6	H	bahasa rupa salah satu unsur identitas suatu b...
1644	4	H	cakup jumlah bahasa saling mirip tutur wilayah...
1645	4	H	empat faktor bahasa melayu angkat jadi bahasa ...
1646	4	H	memang banyak guna bagi besar masyarakat indon...
1647	4	H	bahasa melayu ini akhir jadi bahasa indonesia ...

1648 rows × 3 columns

Figure 4. Preparing Dataset

After successfully importing data from previous studies, the Bag of Words process was carried out and generated 1856 unique tokens.

In the BoW extraction process, the CountVectorizer module performs fitting or training on (value) and (answer) variables by producing 1856 unique token feature extraction from 1648 lines of data or student answer documents. The following is the output of the BoW data extract :

Table 2
Feature Extraction Results

	f1	…	f254	…	f1460	…	f1856
D1	0	…	4	…	1	…	0
…							
D584	0	…	29	…	1	…	0
…							
D1648	0	…	10	…	2	…	0

In the first row that shows f1 – f1856 is a feature word or unique token that appears in the data line of the student answer document which is aimed at columns D1 – D1648. The numbers shown in each document line and feature column are the number of unique token features that appear in each document.

After extracting feature of the text data, the next step is to split the dataset into training data or training and test data. The results/output of the split data can be seen in the following table.

Table 3
Data Split Distribution

Score	Data Train	Data Test
1	3	1
2	154	28
3	19	3
4	146	39
5	98	33
6	227	52
7	129	20
8	257	76
9	109	26
10	104	34
11	15	3
12	41	11
Total	1318	257

In the training data, the total amount of data with a percentage of 80% is 1318 total data. The value eight (8)

e-ISSN: 2723-3944

has the highest frequency with 257 student answer data and the lowest frequency is the value one (1) with 3 student answers. In the test data, the total amount of data with a percentage of 20% is 330 total data. The result is a score of eight (8) as the highest frequency, namely 76 student answer data, and a value of one (1) as the lowest frequency, just one student answer.

After splitting the data, the next step is to calculate regression from test and training values. The results of the regression calculations can be seen in Figure 5.

```
Out[40]: array([ 5.8182701 ,  3.87171718,  6.00831709,  4.02987589,  5.47072239,
                 8.52947396,  7.75016018,  4.77806216,  4.43083663,  4.34993466,
                 5.31732386,  4.16879369,  8.13635597,  8.85914516,  6.35312487,
                 5.63841029,  7.27195282,  7.90028159,  7.69204857,  8.63437742,
                 4.68646533,  8.09281179,  5.71087058,  4.96454493,  7.47324445,
                 4.70479481,  7.27195282,  4.33464211,  4.00625288,  4.68790549,
                 7.68053074,  7.13795288,  6.30033633,  7.9217886 ,  7.27195282,
                 6.24297022,  7.69204857,  6.14975369,  7.69204857,  4.55368849,
                 4.81843285,  4.29811839,  4.07193087,  4.33258693,  5.08159433,
                 9.10006569,  7.28932677,  5.44727367,  7.25371076,  9.10107599,
                 8.01532502,  4.71196901,  7.27195282,  9.89977458,  6.28030624,
                 9.73959013,  4.71196901,  3.8271865 ,  4.32460534,  9.02182255,
```

Figure 5. Results of Regression

```
Out[7]: 'dari dulu bahasa melayu guna bahasa antar indonesia bahasa melayu milik sistem
         sederhana mudah paham ajar suku suku indonesia aku terima bahasa melayu dasar b
         ahasa indonesia dwibahasa melayu milik mampu bahasa budaya'
```

Figure 6. Contents of the First Index

Figure 5 shows the regression results of the selected test data, one of which can be taken as an example is the predicted value of the first index or the first value in the first row is 5.8182701. The first index data (Figure 6) has a score label of 6.

After doing the regression using SVR, the results of the regression are then evaluated for its performance using RMSE, where the smaller the RMSE value, the better the learning model obtained. RMSE result is 1.993.

Table 4
RMSE Comparisson with Previous Research

Method	RMSE
TF-IDF + SVM	2.730
TF-IDF + SVR	2.166
BoW + SVM	1.993

Table 1 shows the classification extraction method of the previous research conducted by Thamrin (2021) TF-IDF + SVM has an RMSE value of 2.730 and is continued by the regression method of research conducted by Verdikha (2021) TF-IDF + SVR has an RMSE value of 2.166. The BoW extraction method and the SVR regression in this study showed a smaller RMSE value compared to the classification and extraction methods in the research that had been carried out by two previous researchers.

5. Conclusion

Based on the RMSE value obtained, the learning model in this study has made comparisons with research methods that have been carried out previously. The Bag of Word extraction method and SVR Regression in this research learning model are proven to be able to reduce the error rate resulting in a smaller RMSE value compared to the method used in previous studies with an RMSE value of 2.73 in Thamrin's research and 2.166 RMSE values in Verdikha's study. This research proves that the research conducted by Thamrin and Verdikha (2021) can be developed even better. Verdikha's research proved that Verdikha's research could reduce RMSE smaller than Thamrin's research by changing the SVM classification method used by Thamrin to the SVR regression method. However, this study proves that changing the Tf-Idf text data extraction to Bag of Word and SVR regression gives a better RMSE value compared to the Tf-IDF used in previous studies.

Continuing on previous research which states that the dataset is classified as inconsistent data in assessing answers and it is hoped that for further research, the data can be reviewed again and produce maximum and better learning modeling results.

Acknowledgements

This work was supported and partially funded by Universitas Muhammadiyah Kalimantan Timur (UMKT) grant no PPI-001.

References

[1] D. A. Perkasa *et al.*, “Sistem Ujian Online Essay Dengan Penilaian Menggunakan Metode Latent Sematic Analysis (Lsa),” *J. Rekayasa dan Manaj. Sist. Inf.*, vol. 1, no. 1, pp. 1–9, 2015, [Online]. Available: http://ejournal.uin-suska.ac.id/index.php/RMSI/article/view/1313

[2] Gunawansyah, R. Rahayu, Nurwathi, B. Sugiarto, and Gunawan, “Automated essay scoring using natural language processing and text mining method,” *Proceeding 14th Int. Conf. Telecommun. Syst. Serv. Appl. TSSA 2020*, pp. 20–23, 2020, doi: 10.1109/TSSA51342.2020.9310845.

[3] D. Yustiana, “Penilaian Otomatis Terhadap Jawaban Esai Pada Soal Berbahasa Indonesia Menggunakan Latent Semantic Analysis,” *Semin. Nas. "Inovasi dalam Desain dan Teknol. - IDeaTech 2015*, pp. 123–130, 2015.

[4] S. Tjandra *et al.*, “IMPLEMENTASI GENERALIZED LATENT SEMANTIC ANALYSIS UNTUK PENILAIAN OTOMATIS JAWABAN ESAI SISWA PADA TINGKAT SEKOLAH MENENGAH,” 2017.

[5] H. Thamrin, N. A. Verdikha, and A. Triyono, “Text Classification and Similarity Algorithms in Essay Grading,” *4th Int. Semin. Res. Inf. Technol. Intelegent Syst.*, 2021.

[6] N. A. Verdikha, H. Thamrin, A. Triyono, M. Fikri, and S. H. Suryawan, “Regression and Oversampling Method for Indonesian Language Automated Essay Scoring,” no. 2, 2021.

[7] A. Sharma and D. B. Jayagopi, “Automated Grading of Handwritten Essays,” *2018 16th Int. Conf. Front. Handwrit. Recognit.*, pp. 279–284, 2018, doi: 10.1109/ICFHR-2018.2018.00056.

[8] C. Lu and M. Cutumisu, “Integrating Deep Learning into An Automated Feedback Generation System for Automated Essay Scoring,” no. Edm, pp. 573–579, 2021.

[9] N. K. Wangsanegara and B. Subaeki, “IMPLEMENTASI NATURAL LANGUAGE PROCESSING DALAM PENGUKURAN KETEPATAN EJAAN YANG DISEMPURNAKAN (EYD) PADA ABSTRAK SKRIPSI

e-ISSN: 2723-3944

MENGGUNAKAN ALGORITMA FUZZY LOGIC Jurusan Teknik Informatika , Fakultas Sains dan Teknologi UIN Sunan Gunung Djati Bandung Ejaan yang," vol. 8, no. 2, 2015.

[10] T. Mardiana, R. D. Nyoto, P. Studi, and T. Informatika, "Kluster Bag-of-Word Menggunakan Weka," vol. 1, no. 1, pp. 1–5, 2015.

[11] R. P. Furi, M. Si, and D. Saepudin, "Prediksi Financial Time Series Menggunakan Independent Component Analysis dan Support Vector Regression Studi Kasus : IHSG dan JII," vol. 2, no. 2, pp. 3608–3618, 2015.

e-ISSN: 2723-3944