Automated Bahasa Indonesia essay evaluation with latent semantic analysis

A Amalia¹, D Gunawan², Y Fithri², I Aulia²

¹Department of Computer Science, Universitas Sumatera Utara, Jl. dr. Mansur No. 9 Kampus USU Medan 20155
²Department of Information Technology, Universitas Sumatera Utara, Jl. dr. Mansur No. 9 Kampus USU Medan 20155

*E-mail: amalia@usu.ac.id

Abstract. Essay Examination is one of many ways to evaluate the learning process of students. The Essay test can measure a student's capability of memory and to develop an idea. With the vast development of Information Technology, an essay test is done in a more sophisticated way with a platform called e-learning. This research implemented an e-learning interface to conduct an essay test, starting with giving out questions, answering it and marking it. The system will generate an automatic mark, using the Latent Semantic Analysis (LSA) method by measuring the relevance of the key answer and the student's answer. Before entering the LSA phase, the answer will go through pre-processing step which is cleansing, case folding, tokenization, stop word, convert negation and stemming. The result of the accuracy of this system compared to teacher's manual assessment is 83.3%. Expectedly, this research will help teachers in making their assessment process more efficient.

1. Introduction

The learning system at higher education in Indonesia mostly uses the Semester Credit System (known in Indonesia as SKS) [1] with 16 discourse classes for each subject. Each discourse is a specific topic that is mutually related to one another. The test evaluation for each topic is needed to know whether learning can be well received or not by students. On the other hand, the evaluation process can also be used as feedback for teachers to improve the quality of the learning process when the results have not been achieved. Test or exam at the end of the course can be a tool to see the achievement of students to absorb the material after the topic is presented. And the test evaluation before a new topic can be done to review student understanding of the previous topic. Teachers may first notice that a student is having difficulty progressing as expected by taking frequent measurements of the student’s classroom performance [2]. This evaluation can improve students' cognitive skills especially for remembering (C1) and understanding (C2). Cognitive ability to remember or C1 is the ability of students to recall relevant knowledge originating from long-term memory, while Understanding or C2 is the ability of students to build an understanding of instructional messages including oral, written, and graphical communication (interpreting, exemplifying, classifying, summarizing, inferring, comparing, explaining)[3]. For time efficiency to carry out the exam can be done using E-learning system. Nowadays many universities including in Indonesia using E-learning in their learning system. E-learning is used as a media to a discussion, teaching material distribution and also exams. Some types of tests that can be done are objective-type question multiple choice questions, short answer, selection/association, hot spots, true/false, visual identification. Most of the on-the-type question exam models are already equipped with automated assessment tools [4] [5]. Unlike the types of multiple choice questions that have been provided with automatic evaluation or grading in the E-learning system, automatic grading for essay questions are still very limited, especially for Bahasa Indonesian essays. Although in some E-Learning Learning Management System (LMS) have a facility to grading short answer questions automatically, the method of assessment is very rigid, because the assessment is
done by looking at the exact similarity of the students' answers with the key. Values are given in a score between 0-100, a value of 0 (zero) is given in an answer that has no resemblance to the key answer, and a value of 100 is given at the same answer with the key [6]. Essay questions are considered better in measuring students' cognitive especially for "remembering" and "understanding" comparing to multiple choice questions. Many researchers agree that essays are the most useful tool to assess learning outcomes, implying the ability to recall, organize and integrate ideas, increasing knowledge in writing and the ability to convey with identification from interpretation and application of data [7]. In carrying out essay examinations, each student is required to answer with his own sentence without being given a choice of answers, so that each student's answer has many variations. Because of this variation, some problems occur such as difficulty in providing the same objectivity for each student's response. An essay is an appropriate evaluation to measure student’s remember ability and show their ideas. So that essay questions that covering all subject material can be an option to facilitate the lecturer in evaluating the learning and teaching process. The problem is manual grading of essays assessment requires a lot of time [4][5]. Considering that evaluations will be conducted at the end of class, it is indeed an obstacle for lecturers if done manually. Unlike the multiple choice which is easier in evaluating based on the answer key, the essay exam requires more effort in the assessment.

The concept of automatically essay evaluation is the comparison similarity of students' answers with the answer key from the lecturer. The main challenge is many variations of student's answer but has the same semantics with the answer key will make difficult to measure this similarity. Some things that must be considered are synonym words and convert negations. Based on this, in this paper, we built an automated system for Automated Essay Evaluation (AEE) which can be integrated into the e-learning system. For teachers, AEE beside as a tool to assess learning outcomes, it also can save time, effort and money without lowering the quality [7]. This study implements the Latent Semantic Analysis (LSA), LSA is one of the methods to analyze semantics [3][5]. We also perform a text mining method, especially for the preprocessing stage. Several related studies show that the proper preprocessing process can improve the optimization of LSA accuracy.

The paper is divided into five sections. Section 2 describes several of the related studies to this research. Section 3 describes the methodology, Section 4 presents result and discussion and Section 5 draws conclusions of this study.

2. Related Works
Automated essay assessment has become essential projects from a long time ago. Many researchers and commercial developers have built this assessment essay system. One of the most famous is Project Essay Grade (PEG). PEG is one of the earliest and longest-lived implementations of automated essay grading. It relies on style analysis of surface linguistic features of a block of text [4].

PEG is not entirely Natural Language Process (NLP) Technique, but it based on the proxes concept. Proxes considered essay length, counts of prepositions, relative pronouns and else to represent the fluency, as an indicator of the complexity of sentence structure. Proxes are calculated from a set of training essays and are then transformed and used in a standard multiple regression along with given human grades for the training essay to estimate the regression coefficients.

Another previous research on automatic essay scoring systems was carried out by [8], this study using the Multivariate Bernoulli Model (MBM) and Bernoulli Model (BM) methods. This method got the accuracy of 80%. Other further research was conducted [9] using the K-Nearest Neighbor algorithm. Each essay was changed into the Vector Space Model (VSM). This study implements the preprocessing process such as stop words removal, feature selection of the essay and the value of each vector is expressed by term frequency-inverse document frequency (TF-IDF). This study implemented cosine in the KNN algorithm to calculate the similarity of an essay and answer key. Experiments on the CET-4 (College English Text) essay in China Learner English Corpus (CLEC) showed accuracy above 76%.
Another Previous AEE is the Intelligent Essay Assessor (IEA) application that applies NLP and text mining [10]. IES based on the Latent Semantic Analysis (LSA) technique. LSA represent documents and their content in a large two-dimensional matrix semantic space [4]. LSA does not consider word orders since the authors claim that this is not the most crucial factor for grabbing the sense of a passage. LSA is a powerful method so that many of the previous studies using this method in designing automated essay grading [11]. This study is devoted to essays evaluation in English, so that there are differences in our research, especially in the preprocessing stage. According to [12] a proper pre-processing process can improve information retrieval, and it could be advantageous for LSA process as well. Research by [13] was observing the impact of weight functions in preprocessing on LSA performance. As in the study [14], this study was used LSA in automatic essay scoring with n-gram features with preprocessing stages are case folding, lexical analysis, and tokenization. Some previous studies in AEE have also been conducted for Bahasa Indonesia as [15] [16]. These studies observe various preprocessing stages. Same as previous studies, in our paper we also implemented preprocessing stages, however this study applied some additional processes in preprocessing, such as checking synonyms and convert negation. The research conducted by [17] declared that check synonym function could increase the optimization for unigram or tokenization process.

3. Methodology

The methodology of this study can be described in Figure 1.

![Figure 1. Methodology](image)

3.1 Data Input
Data input is a lecture’s key answer and students’ answer for each essay question. Both of this inputs will be processed separately in preprocessing. After being treated subsequently, calculation of similarity of this Data input will be done to state the final score.

Example of Lecture’s Key Answer (Q):
Kemampuan mata untuk mencembungkan dan memipihkan lensa mata saat melihat objek. (The ability
of the eye to inflate and flatten the eyepiece when viewing objects.)

Examples of students’ answer
Student 1 (D1) = Kemampuan mata untuk mencembungkan atau memipihkan lensa mata.
Student 2 (D2) = Daya kemampuan mata untuk menebalkan dan memipihkan lensa mata saat melihat objek.
Student 3 (D3) = Kemampuan mata untuk menembalkan dan memipihkan lensa.

3.2 Preprocessing
The purpose of pre-processing stages is to transform the data input into, so that be easier to process in the next step. In this study there are five stages of pre-processing which are cleaning, case folding, tokenizing, removing stopwords and stemming.

3.2.1 Cleaning
This process stage is a process to clean unneeded attributes in Data input such as symbol and punctuations.

3.2.2 Case Folding
Case folding is a process to normalize text into the same form. In this study, the data input is converted into the lowercase letter. This process aims to make all the data input text more likely to match.

3.2.3 Tokenizing
Tokenizing is a process to split a sequence of strings into pieces. The purpose of this stage is breaking the text into a minimal meaningful unit. In this study, we split data input into word by word.

3.2.4 Stopwords Removal
Stopwords removal is a step to exclude some unnecessary words such as common words. Stopwords are the most frequent words in the text but do not have significant meaning. Deleting these words will decrease the size of data input and can improve LSA results [12]. In this study, we use stop words for Bahasa Indonesia from Tala Dictionary [18]. The excerpt of Tala stopwords can be seen in Table 1.

| Table 1. The Excerpt of Tala Stopwords |
|--------------------------|
| ada  berikut  ini  untuk |
| apa  berbagai  tetapi  Maupun |
| atas  berarti  Itu  apakah |
| demikian  bagaikan  oleh  seperti |

3.2.5 Convert Negation
In this stage, a process to identify all negation words. The next process is, to change the polarity of words that have position after negation words into opposite meaning. The examples of negation words in Bahasa Indonesia such as “tidak”, “tak”, “bukan”. The compound words like “tidak bagus (not good)” will be change into “jelek” (bad).
3.2.6 Stemming
The stemming process is a stage to derived words into their base stem or root base. This process aims to reduce various words to be checked and compared to the dictionary. Stemming will increase comparing time process on the next stage. We deployed Nazief-Adriani algorithm [19] to this process. This algorithm that implemented the basic of Bahasa Indonesia morphological rules. This algorithm checked collected allowed affixes and unallowed affixes. This algorithm use basic word Indonesian dictionaries to compare the root word.

3.2.7 Synonym Conversion
In this phase, the system will check and replace all vocabularies from each document that have the same meaning into only one word. We implemented a synonym dictionary in this process. Future work might implement the synonym set based on semantic web [20].

3.3 Latent Semantic Analysis
In this stage, all the documents are presented into a matrix. The dimensions of the matrix have corresponded to the number of terms or vocabularies in the material involved. Every cell of a matrix composed by terms frequency occurrence. A zero value means that the term is not present in the document. The example of this process can be seen in Figure 2.

|     | Q  | D1 | D2 | D3 |
|-----|----|----|----|----|
| Mampu | 1  | 1  | 1  | 1  |
| Mata | 2  | 1  | 2  | 1  |
| Cembung | 1  | 1  | 0  | 0  |
| Pipih | 1  | 1  | 1  | 1  |
| Lensa | 1  | 1  | 1  | 1  |
| Lihat | 1  | 0  | 1  | 0  |
| Objek | 1  | 0  | 1  | 0  |

**Figure 2. Terms Frequency Represent as a Matrix**

LSA (Latent Semantic Analysis) is a technique for analyzing a document to find the meaning or concept of the document by comparing the semantic similarity. The meaning or concept of the words contained in writing will be a reference comparison without looking at the linguistic characteristics of a writing. The LSA method does the mapping of words or documents into a concept space and comparison is done on this space. Concept space or more commonly referred to as latent semantic space is the result of mapping the high dimensional matrix into smaller dimensions. Although in a smaller dimension, the matrix still represents the contents of the entire document. A distinctive feature of LSA is a technique called Singular Value Decomposition (SVD). SVD Is a technique in linear algebra which has many functions in both processing documents and text. SVD is known as a strong technique, concerning solving problems of equations or matrices, both singular and numerically close to singular. SVD is used to decompose the matrix after being weighted to then measure its similarity to the data to be tested [21]. The advantage of SVD, this method can be used in all real matrix sizes (m, n). If A is a real matrix with size m x n and decomposed with Singular Value Decomposition (SVD) to:

\[ A = USV^T \]  

Where
\[ U = m \times m \] orthogonal matrix, which is the AA\(^T\) vector eigen
\[ S = \] diagonal matrix composed of m \times n singular value, which is the root of the eigen value of the
square $AA^T$ and $A^TA$

$V = \text{an orthogonal matrix measuring } n \times n$, which is the eigen vector of $A^TA$

In the string of key answers and in the string of students answer have gone through the preprocessing stage. Then string matching is done to calculate the similarity value between the student's answer and answer keys.

### 3.5 Cosine Similarity

Cosine similarity is a measure of similarity to measure of the angle between the document vector ($D$) and the query vector ($Q$). Each vector represents each word in each document (text) that is compared. The more the same vector document with the query vector is, then the document can be seen to be more suitable with the query. Measurement Cosine Similarity uses equations

$$ Sim(Q, D) = \cos(Q, D) = \frac{Q \cdot D}{||Q|| \cdot ||D||} = \frac{\sum_{i=1}^{n} Q_i D_i}{\sqrt{\sum_{i=1}^{n} (Q_i)^2} \sqrt{\sum_{i=1}^{n} (D_i)^2}} \quad (2) $$

When two documents are identical, the angle is zero degrees ($0^\circ$) and their similarity is one (1); when two documents are not identical at all, the angle is 90 degree ($90^\circ$) and the similarity is zero (0).

### 4. Result and Discussion

This application can be integrated to the E-Learning LMS system. With using this system, the teacher can organize questions, key answer and grading easily. Based on this key answers and this grading value, all the essay answers of the students will be automatically evaluated. The screenshot of this system can be seen in Figure 3.

![Figure 3. Application Appearance](image)

To measure the accuracy of this application, we deployed three scenarios to measure the accuracy of this system. The scenario consists of:

- **Scenario 1**: 100% similarity level, or the student's answer similar to the answer key
- **Scenario 2**: Synonym similarity, or the terms of student’s answer not similar to the answer key, but these terms have the same meaning
Scenario 3: 0% similarity level, or the student’s answer different with the answer key
In this process, each essay answer is given a weight of 5 (five) if the answer is "True." Based on the test results of 20 (twenty) questions with 3 (three) different scenarios can be concluded:

a. The first scenario with 100% similarity level yields a value of 100 where each word and the number of words from the students' answers are the same as the lecturers' answers.

b. In the second scenario, although the system can identify the synonym words, but the difference of the length of the answer key and the student's answer is also influential. The more differences the number of these words, the smaller the similarity level the system will get.

c. In the third scenario, students' answers and lecturers' answers have different compositions. In this evaluation, the system can measure the similarity is 0.

Of the 60 questions and answers tested with all three scenarios, as many as 50 or 83.3% of the evaluation results are as expected. So we can conclude the accuracy of this system is 83.3%.

5. Conclusion
Essay questions are considered better in measuring students' cognitive especially for “remembering” and “understanding” comparing to multiple choice questions. The essay test is the most useful tool to assess learning outcomes, implying the ability to recall, organize and integrate ideas, increasing proficiency in writing and the ability to convey with identification from interpretation and application of data. But the evaluation of the essay questions consuming much time. Therefore, in this paper, we deploy automatic essay grading system with Latent Semantic Analysis to measure the essay test. The methodology of this work contains cleansing data, case folding, tokenization, stop word, convert negation, stemming, LSA and cosine similarity measurement. This application denotes to an essay in Bahasa Indonesia. We implemented additional dictionary to handle the synonym words. The result of the accuracy of this system compared to teachers manual assessment is 83.3%. Expectedly, this research will help teachers in making their assessment process more efficient.

References
[1] Gaus N and Hall D 2016 Performance Indicators in Indonesian Universities: The Perception of Academics High. Educ. Q. 70 127–44
[2] Overton T 2012 Assessing Learners with Special Needs An Applied Approach (Pearson)
[3] Marzano R J 2001 Designing a new taxonomy of educational objectives (Corwin Press)
[4] Valenti S, Neri F and Cucchiarelli A 2003 An Overview of Current Research on Automated Essay Grading J. Inf. Technol. Educ. 2 3–118
[5] Islam M M and Hoque A S M L 2012 Automated Essay Scoring Using Generalized Latent Semantic Analysis
[6] L A R 2017 ANALISIS UJI METODE PEMBOBOTAN KATA TF-IDF-DF DALAM UPAYA MENINGKATKAN HASIL SKORING PADA JAWABAN ESAI BAHASA INDONESIA (Universitas Gadjah Mada)
[7] Zupanc K and Bosnić Z 2017 Automated essay evaluation with semantic analysis Knowledge-Based Syst. 120 118–32
[8] Rudner L M and Liang T 2002 Automated Essay Scoring Using Bayes’ Theorem vol 1
[9] Bin L, Jun L, Jian-Min Y and Qiao-Ming Z 2008 Automated Essay Scoring Using the KNN Algorithm 2008 International Conference on Computer Science and Software Engineering (IEEE) pp 735–8
[10] Foltz P, Laham D and Landauer T 1999 The intelligent essay assessor: Applications to educational technology imej.wfu.edu
[11] Miller T 2003 Essay Assessment with Latent Semantic Analysis J. Educ. Comput. Res. 29 495–512
[12] Wild F, Stahl C, Stermsek G and Neumann G PARAMETERS DRIVING EFFECTIVENESS OF AUTOMATED ESSAY SCORING WITH LSA
[13] Nakov P, Popova A and Mateev P *Weight functions impact on LSA performance*

[14] Citawan R S, Mawardi V C and Mulyawan B *Automatic Essay Scoring in E-learning System Using LSA Method with N-Gram Feature for Bahasa Indonesia*

[15] Agung A and Ratna P *SIMPLE-O, the Essay Grading System for Indonesian Language Using LSA Method with Multi-Level Keywords*

[16] Thamrin H and Wantoro J 2014 *An attempt to create an automatic scoring tool of short text answer in bahasa indonesia* *Int. Conf. Electr. Eng. Comput. Sci. Informatics* 1 96–8

[17] Amalia A, Lydia M S, Fadilla S D, Huda M and Gunawan D *Document Clustering Optimization with Synonym Dictionary Check Function Case Study: Documents in Bahasa Indonesia*

[18] Tala F Z A *Study of Stemming Effects on Information Retrieval in Bahasa Indonesia*

[19] Adriani M, Asian J, Nazief B and Williams H E 2007 *Stemming Indonesian: A Confi x-Stripping Approach* *ACM Trans. Asian Lang. Inf. Process.* 6 1–33

[20] Gunawan D and Amalia A 2017 *The design of lexical database for Indonesian language* *IOP Conf. Ser. Mater. Sci. Eng.* 180 012052

[21] Deerwester S, Dumais S T, Furnas G W and Landauer T K 1998 *Indexing by Latent Semantic Analysis*