Answer Categorization Method Using K-Means for Indonesian Language Automatic Short Answer Grading System Based on Latent Semantic Analysis

Anak Agung Putri Ratna, Naiza Astri Wulandari, Aaliyah Kaltsum, Ihsan Ibrahim, Prima Dewi Purnamasari
Department of Electrical Engineering, Faculty of Engineering
Universitas Indonesia
Depok, Indonesia
ratna@eng.ui.ac.id

***Abstract*— The Automatic Short Answer Grading (Simple-O) has been created for grading short answer with Bahasa Indonesia using K-Means and Latent Semantic Analysis (LSA) method. In this experiment, the text document feature will be extracted using Term Frequency-Inverse Document Frequency (TF-IDF) and then classified using K-Means. From the experiment, 149 documents are expected to be clustered into five classes. The result of the clustering using K-Means is 100% matched with clustering using human rater. The result of grading with LSA is 74%.**

Keywords—essay grading, k-means, latent semantic analysis, answer categorization, tf-idf

I. INTRODUCTION

These days, technology is developing fast in every field, including the education field. E-Learning environment is one of education method that makes it possible to for student and lecturer to do teaching and learning activities without being in the same place. The technology was made to make a process easier. By using E-Learning, will make learning activity more efficient in time, places, and cost.

Besides learning and teaching activity, E-learning also can be used to do an online examination. The online examination usually is done by multiple-choice because it's easy to correct, not like short answer type. Short answer type of test was harder to correct than multiple choice, because every student will give a different answer. Because of this, the short answer type of test usually corrected manually by the lecturer. Correcting short answers manually was not efficient in time and less objective. To resolve that problem, need a system to correct short answer automatically, so the answer can be corrected faster and more objective.

The automatic short answer grading system (SIMPLE-O) has been developed since 2007 by the Department of Electrical Engineering, Universitas Indonesia. At that time, SIMPLE-O was developed using Latent Semantic Analysis to process the text to be a matrix. Since then, SIMPLE-O already developed using many algorithms such as Simple Vector Machine, Winnowing, Levenshtein, etc. Machine learning that used in previous Automatic Essay Grading are also LVQ and NLP [11] .

This paper will discuss developing an automatic short answer grading system (SIMPLE-O) for short answer in Bahasa Indonesia using K-Means algorithm to classify the student answer, and then the answer will be corrected using Latent Semantic Analysis by comparing Frobenius norm of student and lecturer answer.

This paper is organized into sections. After this section, section 2 explained K-Means and LSA. Section 3 proposes the design of automatic essay grading system with K-Means and LSA. Then, Section 4 explained the experiments, results, and analysis. The final section described the conclusions of the research.

II. K-MEANS AND LATENT SEMANTIC ANALYSIS ON AUTOMATIC SHORT ESSAY GRADING WITH ANSWER CATEGORIZATION

Automatic Short Answer Grading (SIMPLE-O) is a system that has been developed by the Department of Electrical Engineering, Universitas Indonesia from 2007 until now. SIMPLE-O was developed using Latent Semantic Analysis (LSA) and Singular Value Decomposition (SVD) to process the text in Bahasa Indonesia, and Japanese [1].

A. Latent Semantic Analysis (LSA)

According to Landauer and Dumais, Latent Semantic Analysis is a method to find a hidden concept in text document [2]. By using it, similarity from a word can be found. LSA method will be extracting and represent contextual meaning from words by using statistics. LSA using Singular Value Decomposition (SVD) to do matrix decomposition, from a matrix that contains words representation. Those matrixes will be an input for the LSA process.

The LSA method is not a traditional natural language or artificial intelligence. LSA doesn't use the humanized dictionary, semantic network, grammar, separator or morphology. LSA only parsing the input to text that defined as unique string character and will be separated into important sample in sentences or paragraph.

B. Term Frequency–Inverse Document Frequency (TF-IDF)

Term Frequency-Inverse Document Frequency is a method of numerical statistics to reflect how important a word to a document. TF-IDF is used to weighting a word and extracting text characteristic [3]. The main idea of TF-IDF is if a word appears frequently in a document than the other document, then the word considered to be a class distinction [9].

TF is for Term Frequency, will measure how frequent a term occurs in a document [4] TF formula is shown in the first equation below (1). IDF is for Inverse Document Frequency, will measure how important a term in a document, IDF formula is shown in the second equation (2):

$$\text{TFij}=\frac{Ni.j}{\sum_{k=0}^{n} Nk.j} \quad (1)$$

$$\text{IDF} = \log\left(\frac{|D|}{|\{dj \in D : ti \in dj\}|}\right) \quad (2)$$

Ni,j is the number of occurrences of the i-th word in document Dj. TF-IDF formula is shown in (3)

$$TF - IDF = TF * IDF = \frac{Ni,j}{\sum_{k=0}^{n} Nk,j} * \log\left(\frac{|D|}{|\{dj \in D : ti \in dj\}|}\right) \quad (3)$$

C. K-Means

K-Means is a clustering algorithm. K-Means basic ideas are separated into two phases [8]. The first phase is, for a given number of cluster k, K number of text will randomly be selected first as the initial cluster center, where the k value is fixed in advanced, the center usually called centroid [7]. The next phase is, each text will be classified according to the distance between each text with the centroid. Euclidean distance is used to determine the distance between data object with the centroid, Euclidean distance function shown in function 5. The nearest data object with the centroid is considered one cluster. The first step is complete and an early grouping is done when all the data objects are included in clusters. Then the process will continue to recalculating the average of the early grouping clusters [8]. This process will continue until the criterion function becomes minimum. Criterion function is shown in (4):

$$E = \sum_{i=1}^{k} \sum_{x \in Ci} |x - xi|^2 \quad (4)$$

Based on (4), x indicates the target object, xi indicates an average of cluster Ci. Then the distance of criterion function is using Euclidean distance that used to determine the nearest distance between data object and centroid. Equation (5) is showing a function of Euclidean distance between vector x and vector y:

$$d(xi, yi) = \left[\sum_{i=1}^{n} (xi - yi)^2\right]^{1/2} \quad (5)$$

The input of K-Means are including the number of clusters, a dataset containing n data objects. The output of K-Means should be a set of k clusters. Complete steps of K-Means are [8]:

1) Select k data object randomly from the dataset as initial cluster centers (centroid).
2) Repeat
3) Calculate distance using Euclidean distance and assign data objects to the nearest cluster.
4) Recalculate the cluster center for each cluster.
5) Stop recalculate when the cluster center not changing

III. Automatic Short Essay Grading System with Answer Categorization Design

The Automatic short answer grading system (SIMPLE-O) designed to do an examination to short answer in Bahasa Indonesia. Student short answer will be matched with answer key made by the lecturer. This system using two algorithms that is K-Means and Latent Semantic Analysis (LSA). Student short answer will be input to the system as a text document. In order to make the text document possible to be processed in the system, the document needs to pass through pre-processing. The next process is classifying data using K-Means. The classifying process is done to make sure that all the student answer that will be examined is in relevant, so the answer that is not relevant wouldn't be examined. After the classification process, the document will be examined using LSA. This process will find the Frobenius norm of both student and lecturer short answer, the more similar the Frobenius norm is, the better grade the student will get. The flowchart of the process shown in Figure 1.

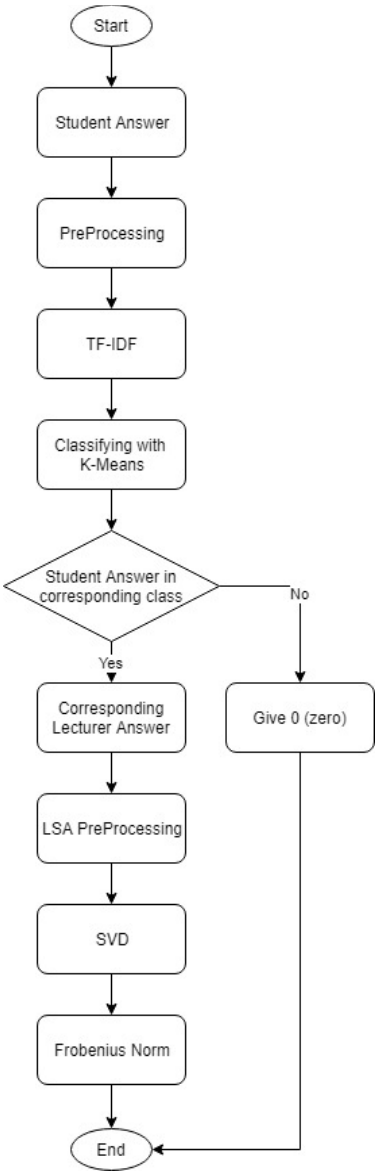

Fig. 1. SIMPLE-O flowchart

A. Pre-processing

This step will process the text document to be a term that can be used for the next process. There are some steps in pre-processing, that is stemming, stopword removal, so there is only important words left that can represent the meaning of the text.

Stopwords removal is done because stopwords don't have significant semantic meaning. Words that belong to stopwords is a preposition, conjunction, and other words that don't represent the sentences. With stopwords, existence will make LSA biased in determining sentence context.

The stemming process will make the word to basic word, so only basic word that will proceed. Stopwords and stemming process are done using a library on python Sastrawi.py. This library works for text in Bahasa Indonesia.

After cleaning the sentences, the next process is TF-IDF. TF-IDF will give each word a weight by measuring how much the term used and how important the term in the sentences. By using TF-IDF, text document can proceed to the next step, that is K-Means because K-Means only process numerical data.

B. K-Means

After getting the result from TF-IDF, then it will become the input for the classifying process using K-Means. K-Means will randomly select the cluster center from each term weight from TF-IDF. The number of the cluster need to be specified first [10]. Then Euclidean distance between weight from TF-IDF and cluster center will be calculated. After getting the Euclidean distance, and get each data in their cluster, the cluster center will be recalculated. The recalculate process will stop when the cluster center is no longer changing.

The classified data will get a label of the cluster. The cluster label will be used for the next process, where student answer and lecturer answer will be matched using LSA. The detailed process of K-Means clustering shown in Figure 2.

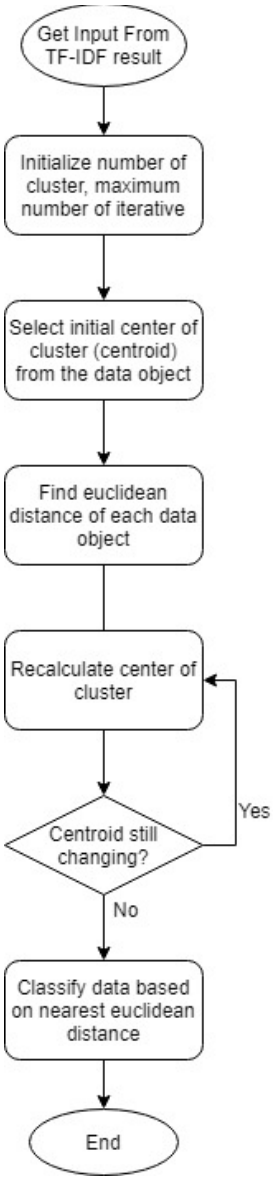

Fig. 2. K-Means algorithm flowchart

C. Latent Semantic Analysis (LSA)

After being classified, the short answer that is not eliminated will go to the next process, Latent Semantic Analysis (LSA). LSA will use student answers and lecturer answers as an input, then the input will be represented in Term Document Matrixes (TDM) then the TDM will be reduced using Singular Value Decomposition (SVD).

SVD will reduce the dimension of the matrix so it will be less complex to proceed. Then Frobenius norm will be used to calculate the vector angle between the student short answer and the lecturer short answer. The similarity between two short answers will be shown in Frobenius norm result. The bigger Frobenius norm is, then the more similar the document is.

IV. Experiments and Analysis

This paper using 148 essay answer data from 37 students including 8 dummy data which is an irrelevant answer. The data will pass through pre-processing to make the text data into numerical data. Using library Sastrawi.py to stemming and remove stopwords from the data which will be used. After stemming and removing stopwords from the data, TF-IDF will be used to give weight to each term. Table 1 showing a few of terms that TF-IDF manages to get.

Each term frequency on each document will be calculated. Document column showing the number of strings that used as input. Terms like “adik”, “ayam”, “bakar”, “berwarna”, “biru”, is a few of terms that extracted using TF-IDF. When the terms that don’t exist in the document, will be given zero, and the terms that exist in the document will get the weight. The more frequent the term in the document exists, then the weight will get bigger.

The next process is classifying the seven-document using K-Means algorithm. Library used is KMeans.py from Sklearn.cluster. The data will be classified into 5 class, which is 4 class for the number of exam question and 1 class for irrelevant student answer. K-Means process will be repeated with the maximum number of iterative is 300 by using the library.

Top terms per cluster can be known from cluster centers that have been used. Table 1 showing top terms per cluster obtained from K-Means process.

TABLE I. Top Terms in Each Cluster

Cluster	Top terms per cluster
Cluster 0	orang fisikawan matematikawan semangat
Cluster 1	server client jaring data minta
Cluster 2	hibernate komputer mati kembali pc
Cluster 3	unit komputer von neumann alu
Cluster 4	boost turbo cepat 4 komputer

Top terms per cluster that shown in table 2, is the term that frequently shown in each document and has similarity with another document. The document that belongs to each cluster can be obtained by using top terms per cluster as reference. If the document having the same keyword (top terms per cluster), then it can be clustered as 1 class. From top terms per cluster result, we can see that cluster 0 is a cluster for irrelevant student answers. The result of the classification process is shown in Table 2.

TABLE II. CLASSIFIED DATA

Question Number	Class Label	Total Document
1	3	35
2	2	35
3	1	35
4	4	35
Irrelevant Answer	0	8

The result of classified data using K-Means and by human rater matched 100%. After being classified, the data will enter the condition if the answer is in the corresponding cluster then the answer will be assessed with corresponding lecturer answer. If the data not in the corresponding cluster, then the data will be graded with zero (0) without even entering the LSA process. When entering the LSA process, pre-processing will be done again for student answer and lecturer answer. The result of the LSA process after K-Means is shown in and Figure 3.

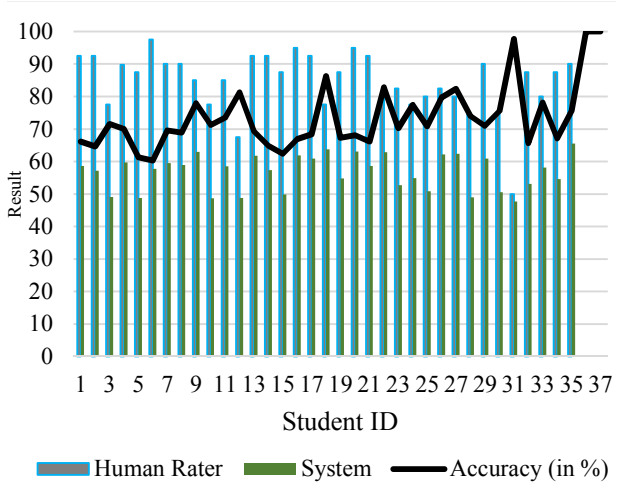

Fig. 3. Result comparison between human rater and system with K-Means answer categorization

From Figure 3, the lowest accuracy obtained by student 6 with 60% accuracy. The highest accuracy obtained by student 36 and student 37 which is a dummy data with irrelevant answers, so the answer from student 36 and 37 are classified to class with label 0 and straight up getting 0 without entering LSA process. The accuracy average is so low because by analysing student answer and lecturer answer, students are using a different word from lecturer to represent the meaning of the answer. So the system considers that the answer is wrong even though the student answer actually true.

Based on the analysis, system of automated short essay grading can get a better result if lecturer give more words variations so even though the student using different words to represent the answer, the system still consider the student answer as a true answer.

Besides using K-Means to classifying the student answers, this paper also compares it with the system that doesn't classify the student answers with K-Means and directly grading all student answers with LSA. The result of SIMPLE-O that doesn't use K-Means is shown in and Figure 4.

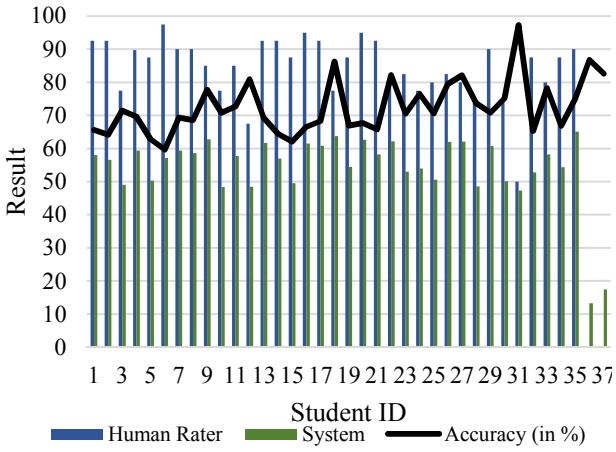

Fig. 4. Result comparison between human rater and system without K-Means answer categorization

From Figure 4, SIMPLE-O with K-Means show better result than SIMPLE-O without K-Means, because the irrelevant answers still graded in SIMPLE-O without K-Means. To see the comparation between SIMPLE-O with K-Means and SIMPLE-O without K-Means is shown in Figure 5.

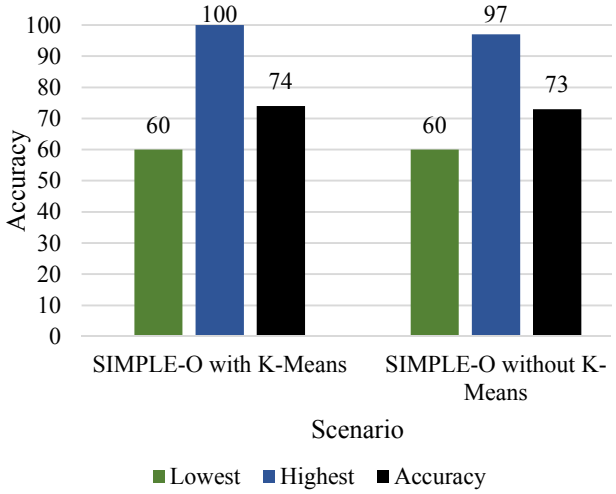

Fig. 5. Results comparison between systems with and without K-Means

V. CONCLUSIONS

From the experiment, SIMPLE-O with K-Means shows the better accuracy than LSA without K-Means. Result of SIMPLE-O using K-Means and LSA is 74% accuracy that have 1% difference with SIMPLE-O without using K-Means which is 73%. SIMPLE-O can get better result if lecturer provide more words variation for grading students essay.

ACKNOWLEDGMENT

This research could not be held and achieved without the full support from Universitas Indonesia and Ministry of Research, Technology, and Higher Education under the grant of PITTA-B 2019 (Publikasi Internasional Terindeks untuk Tugas Akhir Mahasiswa 2019) with contract number NKB-0702/UN2.R3.1/HKP.05.00/2019.

REFERENCES

[1] Agung, A., Ratna, P., Luhurkinanti, D. L., Ibrahim, I., Husna, D., & Purnamasari, P. D. (2018). Automatic Essay Grading System for Japanese Language Examination Using Winnowing Algorithm. 2018 International Seminar on Application for Technology of Information and Communication, 565–569.

[2] "Latent Semantic Analysis Tutorial, 1–7" [Online]. Available:http://bi.snu.ac.kr/Publications/Theses/ShinDH_MS00.pdf

[3] Azara, M. (1989). Arabic Text Classification Using Learning Vector Quantization. 2012 8th International Conference on Informatics and Systems (INFOS), NLP-39-NLP-43.

[4] Pilevar, M. T., Feili, H., & Soltani, M. (2009). Classification of Persian textual documents using Learning Vector Quantization. 2009 International Conference on Natural Language Processing and Knowledge Engineering, 1–6. https://doi.org/10.1109/NLPKE.2009.5313761

[5] Elsayad, A. M. (n.d.). Classification of Breast Cancer Database Using Learning Vector Quantization Neural Network, (December 2006), 1–9.

[6] Dike, H. U., & Zhou, Y. (2018). Unsupervised Learning Based On Artificial Neural Network : A Review. 2018 IEEE International Conference on Cyborg and Bionic Systems (CBS), 322–327. https://doi.org/10.1109/CBS.2018.8612259

[7] Wu, G., & Lin, H. (2015). An Improved K-means Algorithm for Document Clustering. 2015 International Conference on Computer Science and Mechanical Automation (CSMA), 65–69. https://doi.org/10.1109/CSMA.2015.20

[8] Na, S., Xumin, L., & Yong, G. (2010). Research on k-means Clustering Algorithm: An Improved k-means Clustering Algorithm. 2010 Third International Symposium on Intelligent Information Technology and Security Informatics, 63–67. https://doi.org/10.1109/IITSI.2010.74

[9] Liu, C., Sheng, Y., Wei, Z., & Yang, Y. (2018). Research of Text Classification Based on Improved TF-IDF Algorithm. 2018 IEEE International Conference of Intelligent Robotic and Control Engineering (IRCE), (2), 218–222.

[10] Xiong, C., & Lv, K. (2016). An Improved K-means text clustering algorithm By Optimizing initial cluster centers, 272–275. https://doi.org/10.1109/CCBD.2016.059

[11] A. Shehab, M. Elhoseny and A. E. Hassanien, "A hybrid scheme for Automated Essay Grading based on LVQ and NLP techniques," 2016 12th International Computer Engineering Conference (ICENCO), Cairo, 2016, pp. 65-70. doi:10.1109/ICENCO.2016.7856447.