Detector Similarity Answers Between Students on Essay Digital Exam System

Regi Ismayana Pratama[1], Munir[2], Rani Megasari[3]
{regi@student.upi.edu[1], munir@upi.edu[2], megasari@upi.edu[3]}

Universitas Pendidikan Indonesia, Bandung, Indonesia[1,2,3]

Abstract. This research is motivated by the problem of students who cheat while exam. Cheating can interfere in the evaluation process. A system was made for essay type questions because exams that use computers/smartphones are still many using multiple-choice questions. The method used for essay scoring is term frequency, n-gram, and cosine similarity. The concept is comparing keyword answers as query and student answer as a document then get a score of student answer. This method can also be used in another way, which is comparing answers between students where the result is a score of similarity student answers. Student answers that have 75% similarity will be considered by the system as cheating which is expected to help teachers for evaluating exams or learning.

Keywords: Evaluation, Term Frequency, N-Gram, Cosine Similarity, Exam, Student Cheat

1 Introduction

Cheating is common behavior on students in Indonesia that can be found from elementary school, junior high school, senior high school, until college/university level. One of the research with the subject of 239 students at Universitas Islam Nasional Walisongo, 98% of the subjects shows that cheated once [1]. According to Anderman and Murdock, cheating can reduce the function of the use of data assessment as an indicator of student learning achievement and also a source of reference for teachers in conducting actions and giving feedback [2].

Assigning assignments or exams is a way or a teacher's effort to evaluate learning. Learning evaluation is activity or progress that systematic, sustainable, and comprehensive in order to control, guaranteeing and quality determination (value and meaning) of learning for various learning components, based on certain considerations and criteria, as teacher accountability in conducting learning [3]. Every learning needs evaluation to know the level of thinking and understanding of students toward learning theory/material, either in cognitive, psychomotor, or affective aspects [4].

To evaluate learning, the teacher usually uses an evaluation tool. The evaluation tool is something used in the evaluation process or system. Evaluation tools can be known as an evaluation instrument. Many evaluation tools are used in evaluation activities such as oral tests, writing tests, observation, and interviews. Writing tests can be differentiated into two forms: an objective test and subjective test (essay test). An objective test is testing that can be assessed objectively. It is intended to overcoming weaknesses from subjective tests. An

MSCEIS 2019, October 12, Bandung, Indonesia

DOI 10.4108/eai.12-10-2019.2296348

example of the objective test is a true-false test, multiple-choice test, matching test, field test. While the subjective tests are generally essay-shaped form. An essay-shaped test is a type of learning progress test that requires discussion or description of words [5].

Writing test is often used for evaluation activities as well as daily exams, midterm exams, final exam, and or national exams. Currently in Indonesia, the evaluation tool for writing tests usually uses paper media. But now there is already implemented with digital media or computer-based media. The example of this digital test is a national exam called Ujian Nasional Berbasis Komputer (UNBK). In Bandung, this digital test already implemented on the school scale, one of them is Sekolah Menengah Kejuruan Pekerjaam Umum Negeri Kota Bandung that called Ujian Dalam Jaringan (UDJ).

Generally, in the exam using paper media, there is always a subjective test. however, the system or application used in UNBK or UDJ, the question is used only in the form of multiple choice. Using technology, an essay or subjective test can be implemented in line with the number of system developments for the essay test. The goal of the development of the essay test is to reduce the subjectivity of judgment that can not be done by human assessment.

One of the methods often used in the automatic essay scoring is term frequency and cosine similarity. The usual way to check the similarity between the two documents is to check the words contained in the two documents by calculating the frequency of the words on the two documents, one of the methods is term frequency [6]. The results of the frequency calculation will resulting in a vector space model (VSM). Similarity measurement often used for document retrieval and information retrieval system by calculating the vector of vector space model on each document and cosine similarity is one measurement to compare similarities between two vectors [7]. Many research shows that the term frequency and cosine similarity method can be used as a way of scoring essay questions. However, researchers will try to implement the term frequency, N-gram, and cosine similarity methods. N-gram is used to view or pay attention to the pharse of a document or student answers. N-grams are the order of n items of the given order. The Item can be a phoneme, syllable, letter, or other, which corresponds to the application [8]. The addition of n-gram to the method will be implemented to determine the similarity of answers between students. So in addition to being able to score students' answers from teacher answers, the system can tell/suggest teachers that there are students' answers similar to other students' answers.

2 System Design

Figure 1 is a flowchart of a system for assessing the similarity of the student‘s answer. In this system, there is a function to carry out the exam. From these results, the answers of students have been saved in the database. Each student's answers are processed using text preprocessing to get the basic words of each student to answer. After that, it is weighted on each of the words in the student's answers with the other student's answers in one class. Then an assessment of similarities using cosine similarity. This assessment will get a value that ranges from 0-1. The higher the value, the higher the similarity between one of the student's answers and the other student's answers. If the similarity value is more than the specified similarity condition, it will be marked and entered into the database.

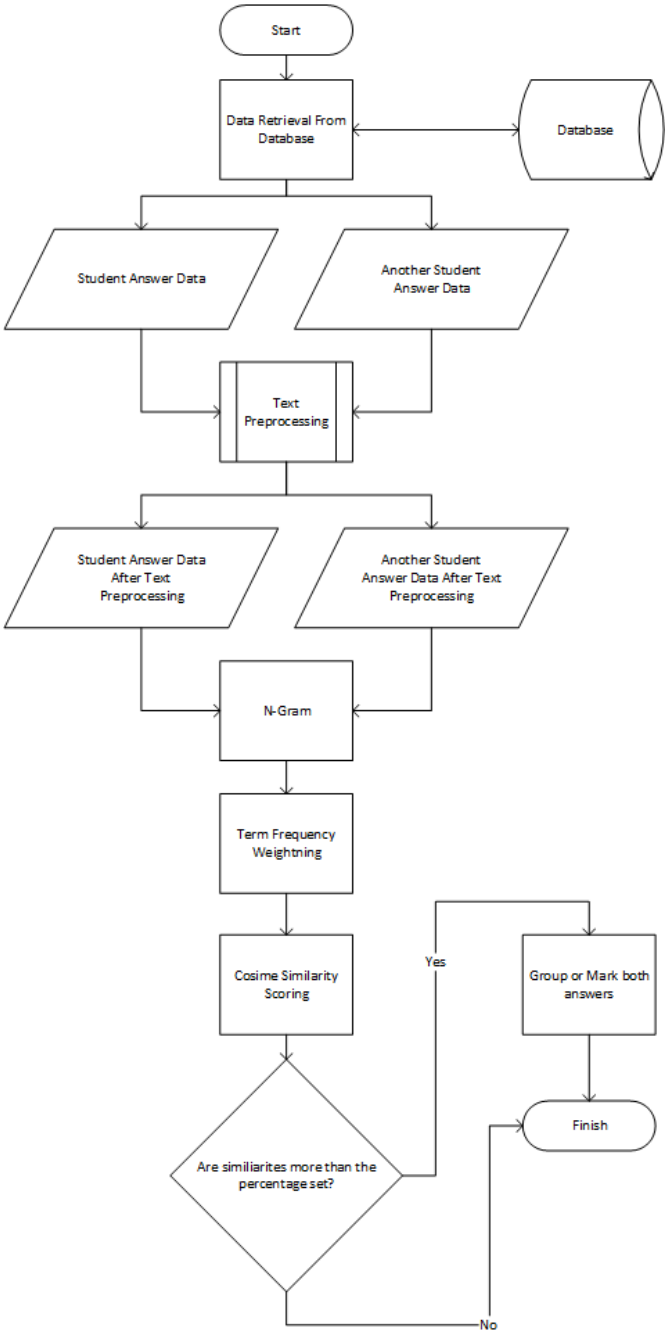

Fig. 1. Flowchart of similarity in students answer.

To get the basic word, a text preprocessing process is carried out with the details depicted in **Figure 2**.

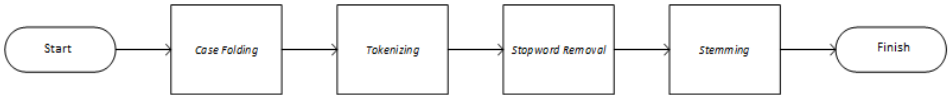

Fig. 2. Flowchart of Text Preprocessing.

2.1 Text Processing

Case Folding

Case folding is the process of equating cases in a document. This is done to equate all consistency in the use of cases such as capital letters. Therefore it takes case-folding to change the whole text to a standard shape, which is used lowercase letters as shown as **Figure 3**.

Initial Sentence:
Kegiatan untuk mempertahankan pendapat dengan disertai argumen yang mendukung.
Case Folding Result:
kegiatan untuk mempertahankan pendapat dengan disertai argumen yang mendukung.

Fig. 3. Case folding process.

Tokenizing

According to Habert and Friends, Tokenization represents the first work done in document processing [9]. Tokenization can be defined as to separate a stream of characters into words. Such as:

- Removing useless tags that remain from the letter drafting information in the archive of the newspaper;
- Retrieving "non-textual" items such as horizontal lines and page-break marks in HTML documents
- Removes parts that are not included in natural languages such as mathematical or chemical formulas as shown as **Figure 4**.

Initial Sentence:
kegiatan untuk mempertahankan pendapat dengan disertai argumen yang mendukung.
Tokenizing Result:
kegiatan untuk mempertahankan pendapat dengan disertai argumen yang mendukung

Fig. 4. Tokenizing process.

Stopword Remover

According to Lo and his friends, by definition, recognized is a meaningless word that has weak discrimination power [10]. Stopword is a non-keyword term or a word that is not meant so it can be removed or deleted in order to reduce the word to be processed. An example word

of stopword in Indonesian is conjunctions of words such as “yang”, “di”, “ke”, etc. Example of Stopword Remover is depicted in **Figure 5**.

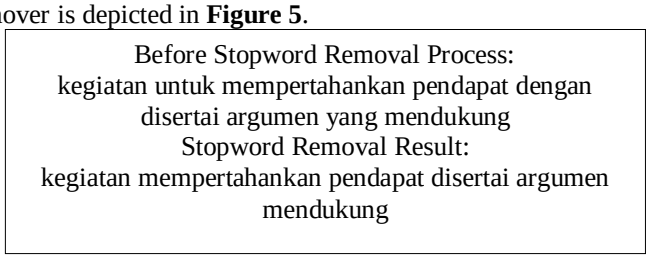

Fig. 5. Stopword remover process.

Stemming

Words often have meaning. There are many forms of words that can affect the meaning of the word, in Bahasa Indonesia is known by the suffix, which is often called morphology. For example, stemming is a method/way of knowing the basic word of a word that is in return. Stemming is usually done by removing suffix and prefix from the word index [11]. Bahasa Indonesia and English have different morphologies. For example, in English, it is necessary to remove the prefixes and suffixes, but in Bahasa Indonesia must be added to remove the confixes [12]. Examples of a stemming algorithm for Bahasa Indonesia is the Porter algorithm and the Nazief-Adriani algorithm.

Stemming in this study uses libraries from Sastrawi. In the Sastrawi library, the main function is the library for stemming. Some of the techniques or algorithms used in this library are:

- Nazief and Adriani Algorithm
- Effective Techniques for Indonesian Text Retrieval
- Enhanced Confix Stripping Stemmer and Ants Algorithm for Classifying News Document in Indonesian Language
- Implementasi Modifikasi Enhanced Confix Stripping Stemmer Untuk Bahasa Indonesia dengan Metode Corpus Based Stemming

Example of this stemming process is depicted in **Figure 6**.

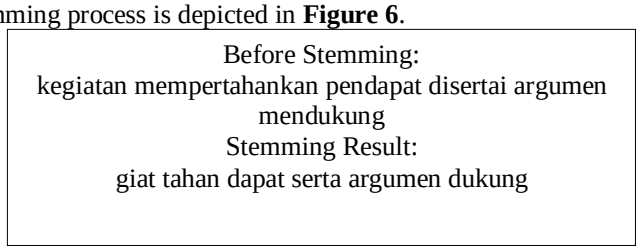

Fig. 6. Stemming process.

2.2 Term Frequency

Term frequency (TF) is the simplest measurement in the weighting method. In this method, each term is assumed to have a proportion of interests according to the number of occurrences in the text (document). Term frequency can improve recall value in information retrieval, but not always improve the precision value [13].

1. Binary TF, rated or weighted with value one if the word is in the document and zero value if the word does not exist in the document.

2. Raw TF, rated or weighted by a value based on the appearance/occurrence of a word in the document. For example, if the word appears three times the value will be three.
3. Logarithmic TF, rated or weighted to avoid the dominance of documents that contain little words in the query, but have a high frequency.

$$TF = \begin{cases} 1 + log(TF)\,,\ tf_{t,d} > 0 \\ 0, tf_{t,d} = 0 \end{cases} \quad (1)$$

4. Normalization TF, use a comparison between the frequency of a word and the total number of words in the document.

$$TF = 0.5 + 0.5 \times \left(\frac{TF}{\max TF}\right) \quad (2)$$

2.3 N-Gram

N-Gram is a sequence n word in a sentence or document. Each language used by humans has some words that appear more often than other words, so the frequency of appearance N-Gram differs from one language to another. Because of the N-Gram characteristics based on the part of a string, it also causes differences in the frequency of N-Gram profiles that the documents have in different languages. The profile is that describes the character of a document formed from N-Gram that is owned by a language document. In generating a frequency profile the N-Gram system reads the document into input then calculates the number of occurrences of each N-Gram in the document [14].

2.4 Cosine Similarity

Similarities between documents are calculated using a function of the similarity measure. The larger the result of the similarity function, then the two objects evaluated are increasingly similar, and vice versa. The quality of the documents obtained depends heavily on the similarity function used [15].

Cosine Similarity is a method used to compute similarity (the level of similarity) between two objects. In general, the calculation method is based on vector space similarity measure. This cosine similarity method calculates the similarity between two objects (e.g. D1 and D2) which are expressed in two vector pieces using the keywords of a document as a measure [16].

This measure allows the alignment of documents according to their similarity (relevance) to the query. The size of the cosine similarity calculates the angular cosine value between two vectors. If there are two vector document (d_j) and query (q) And the term extracted from the document location then the cosine value between d_j and q defined as follows:

$$CosSim(q, d) = \frac{\sum_{j=1}^{t}(q_i * d_i)}{\sqrt{\sum_{j=1}^{t}(q_{ij})^2 * \sum_{j=1}^{t}(d_{ij})^2}} \quad (3)$$

Notes: q = Vector query, d = Vector document, t = Number of vector components d and q, d_i = Vector component d, q_i = Vector component q

3 Experimental Result

An assessment of the student's answers with a specified percentage is 75%. Table 1 shows that there were 50 students answers who concluded similarly to 75% are said to be cheating, 45 students answers are considered cheating by the teacher and 5 answers students are deemed not cheating by the teacher.

Table 1. Results of teacher assessment of system assessment in the similarity between students answers.

Question Number	Group Answers	The similarities of students answers who are considered cheating
2	Group 1	3 of 3 students answers
3	Group 1	2 of 2 students answers
	Group 2	2 of 2 students answers
6	Group 1	4 of 4 students answers
	Group 2	2 of 2 students answers
	Group 3	0 of 2 students answers
	Group 4	7 of 7 students answers
	Group 5	0 of 2 students answers
	Group 6	1 of 2 students answers
7	Group 1	3 of 3 students answers
	Group 2	2 of 2 students answers
	Group 3	2 of 2 students answers
	Group 4	7 of 7 students answers
	Group 5	2 of 2 students answers
9	Group 1	2 of 2 students answers
10	Group 1	2 of 2 students answers
	Group 2	2 of 2 students answers
	Group 3	2 of 2 students answers

From **Figure 7**, can be seen that 90% of the results of the assessment or recommendation of the similarity between students answers is considered cheating so that it can be said that the system using the term frequency, N-gram (trigram) method, and cosine similarity can be said is good to assess similarities between student answers that can advise which students are cheating and which are not, so system able to support learning evaluation.

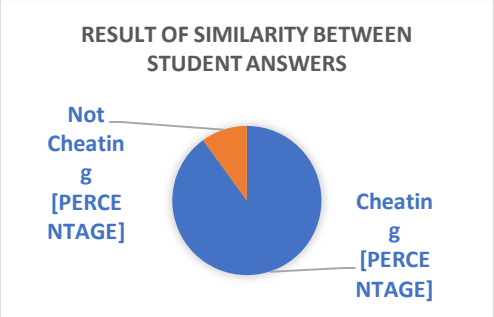

Fig. 7. Chart Result of Similarity Between Student Answers.

4 Conclusion

Based on the research that has been done, can be concluded as follows:

1) This digital exam system has input in the form of question data teacher answer key, and student answers data. The data is processed using preprocessing in the form of the use of Sastrawi library which includes case folding, tokenizing, stopword removal, and stemming. There is an assessment of the similarity between students' answers using term frequency, n-gram, and cosine similarity methods. The output in the form of recommending student answers that are similar to other students who are considered cheating.
2) The implementation of term frequency, N-gram, and cosine similarity methods for the recommendation of similarity between students' answers indicating cheating, with a value of 75% similarity, resulted in a recommendation of 50 similar students answers. Out of 50 students, answers are considered to be cheating or advised by the system, there are 5 students answers considered by the teacher not cheating. When presented, it is 90% of the recommendations by the system in accordance with the teacher's judgment on the cheating answer. It can be said that the system is able to support teachers in evaluating exams or learning.

References

[1] Hadjar, I.: The Effect of Religiosity and Perception on Academic Cheating among Muslim Students in Indonesia," Journal of Education and Human Development. pp. 139-147 (2017)
[2] Cahyo, S. D.: Faktor-faktor yang Mempengaruhi Perilaku Menyontek Pada Pelajar dan Mahasiswa di Jakarta," JP3I. pp. 87-96 (2017)
[3] Arifin, Z.: EVALUASI PEMBELAJARAN, Jakarta: Direktorat Jendral Pendidikan Islam Kementrian Agama RI, (2012)
[4] Ratnawulan, E., Rusdiana, A.: Evaluasi Pembelajaran, Bandung: Pustaka Setia, (2014)
[5] Arikunto, S.: Dasar-Dasar Evaluasi Pendidikan, Jakarta: Bumi Aksara, (2013)
[6] Cavalcanti, E. R., Pires, C. E., Cavalcanti, E. P., & Pires, V. F.: Detection and Evaluation of Cheating on College," Informatics in Education. pp. 169-190 (2012)

[7] Ali, I., Asif, M., Shahbaz, M., Khalid, A., Rehman, M., & Guergachi, A.: Text Catgorization Approach for Secure Design Pattern Selection Using Software Requirement Specification," IEEE Access. pp. 73928-23939 (2018)
[8] Oduntan, O. E., Adeyanju, I. A., Olabiyisi, S. O., & Omidiora, E. O.: Evaluation of N-gram Text Repretentations for Automated Essay-Type Grading System," International Journal of Applied Information System. pp. 25-31 (2015)
[9] Habert, B., Adda, G., Adda-Decker, M., de Marëuil, P. B., Ferrari, S., Ferret, O., Illouz, G., & Paroubek, P.: Towards Tokenization Evaluation," Proceedings of LREC. pp. 427-431 (1998)
[10] Saif, H., Fernández, M., He, Y., & Alani, H.: On stopwords, filtering and data sparsity for sentiment analysis of Twitter.," rev. LREC 2014, Ninth International Conference on Language Resources and Evaluation, Reykjavik, (2014)
[11] Jivani, A. G.: A Comparative Study of Stemming Algorithms," International Journal of Computer Technology and Applications. pp. 1930-1938 (2011)
[12] Widjaja, M., & Hansun, S.: IMPLEMENTATION OF PORTER'S MODIFIED STEMMING ALGORITHM IN AN INDONESIAN WORD ERROR DETECTION PLUGIN APPLICATION," International Journal of Technology. pp. 139-150 (2015)
[13] Tokunaga, T., & Makoto, I.: Text Categorization Based On Weighted Inverse Document Frequency, Tokyo, Japan Tokyo Institute of Technology (1994)
[14] Zaman, B., Hariyanti, E., & Purwanti, E.: Sistem Deteksi Bahasa pada Dokumen menggunakan N-Gram," JURNAL MULTINETICS VOL. 1 NO. 2 NOVEMBER 2015. pp. 21-26 (2015)
[15] Herwijayanti, B., Ratnawati, D. E., & Muflikhah, L.: Klasifikasi Berita Online dengan menggunakan Pembobotan TF-IDF dan Cosine Similarity," Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer e-ISSN: 2548-964X. pp. 306-312, (2018)
[16] Nurdiana, O., Jumadi, J., & Nursantika, D.: Perbandingan Metode Cosine Similarity dengan Metode Jaccard Similarity pada Aplikasi Pencarian Terjemah Al-Quran dalam Bahasa Indonesia," JOIN | Volume I No. 1 | Juni 2016 ISSN 2527-9165. pp. 59-63 (2016)