International Journal of Informatics and Information Systems ISSN 2579-7069
Vol. 5, No. 2, March 2022, pp. 76-83

Implementation of the Jaccard Similarity Algorithm on Answer Type Description

Riyanto [1,*] and Abdul Azis [2]

Universitas Amikom Purwokerto, Indonesia
[1] riyanto@amikompurwokerto.ac.id [*]; [2] abdazis9@amikompurwokerto.ac.id;
* corresponding author

(Received: December 8, 2021 Revised: January 22, 2022 Accepted: February 10, 2022, Available online: March 22, 2022)

Abstract

This study aims to measure the similarity of the answers to the description by using alternative answers as reference answers provided by the lecturer with a view to overcoming the diversity of student answers. This research focuses more on Indonesian language questions and answers by combining the jaccard similarity algorithm and keyword similarity. The results obtained indicate that by adding alternative reference answers, it can increase the correlation value to 0.78% and reduce MAE to 0.55. Likewise, after combining the jaccard similarity algorithm and keyword similarity, the correlation value increased to 0.78% and MAE decreased to 0.49.

Keywords: Text Mining; Problem Description; Jaccard Similarity; Data Preprocessing

1. Introduction

Along with the development of technology, the current teaching and learning process using E-learning is increasing, generally due to the Covid-19 pandemic. E-learning is one of the learning methods in which the learning process, teaching process and even the assessment process are carried out electronically via the internet [1]. By implementing E-learning, the assessment of learning outcomes can be done automatically using the Automatic Grading System. This system has advantages such as being able to score answers quickly and objectively [2].

In higher education institutions, the majority of lecturers ask questions in the form of descriptions. Description questions are a form of question where answer choices are not provided so that students have to answer with sentences [1]. Description questions are the right method for assessing the results of learning activities, because descriptive questions involve students' ability to remember and express their ideas [3]. The problem in the assessment of the description is about subjectivity, the assessment between one lecturer and another may be different. Another problem is the possibility of lecturers having errors in research such as the same student answers but having different scores. Lecturers take a long time to correct the exam. Questions and answers to descriptions require further natural language processing. Therefore, an automatic description assessment system is needed in e-learning to make it easier for teachers to check the results of student description answers. Many researches on automatic assessment have been carried out before but the majority of the datasets used are questions and answers in English. This study aims to measure the similarity of the description answers by developing a method that can be used in any domain without using a word semantic device. In addition, this study also uses alternative answers as reference answers provided by the lecturer with the aim of overcoming the diversity of student answers. This research focuses more on Indonesian language questions and answers by combining the jaccard similarity and keyword similarity methods.

2. Literature Review

2.1. Text Mining

Text mining is an interdisciplinary field that refers to information retrieval, data mining, machine learning, statistics, and computational linguistics [4]. In general, the concept of text mining work is similar to data mining, namely predictive mining and descriptive mining. Text mining extracts a meaningful numeric index from the text and then the information contained in the text will be accessed using various data mining algorithms [5][6].

2.2. Essay Questions

The description question is an evaluation test where the form of the question is structured and does not provide answer choices. Examinees organize their own answers to each question so that answers can vary greatly according to the thoughts of each examinee [7]. Problem descriptions are used to measure higher abilities in cognitive aspects. Problem descriptions are divided into 2 types [8][9], namely Limited or structured descriptions (Restricted Response) and unlimited or free descriptions (Extended Response). The difference between these two types of essay questions is on the basis of the amount of freedom given to test takers to organize, write and express their thoughts, level of understanding of the subject matter and ideas. The description questions are the teacher's choice in evaluating the ability level of their students, even though in reality it is not easy to give an objective assessment of each student's answers. One of the advantages of essay questions is that they can encourage and familiarize students with the courage to express their opinions using their own sentence structure and language style. While the weakness is how to correct the answers to the description questions is quite difficult and takes a long time. This is because the answers to the description test questions can be long and wide and varied, so correcting the answers takes energy, thought, and time [10][11].

2.3. Jaccard Similarity

Jaccard similarity is an algorithm that has a function to compare documents and calculate the similarity value (similarity) of two objects or documents [12][13][14]. Jaccard similarity is defined as the intersection size divided by the union size of the set of samples. Jaccard similarity can be formulated in equation 1 below:

$$Sim_{Jaccard} = \frac{|A \cap B|}{|A \cup B|} \quad (1)$$

Where:

A = Document 1

B = Document 2

2.4. Preprocessing data

In text data, it is necessary to preprocess data, which is to convert text data into numeric data that can be processed. This stage is a very important stage before starting the automatic assessment calculation process because this process can affect the accuracy of the assessment. The preprocessing stage of the text consists of 4 stages [15] [6], namely:

2.4.1. Case folding

Case folding aims to convert the entire text into lowercase letters [16]. In the folding case, only letters 'a' to 'z' are accepted. Characters other than letters are omitted and are considered as delimiters.

2.4.2. Tokenizing

Tokenizing is the process of dividing text from sentences or paragraphs into certain parts [17]. The separator between tokens is a space, enter, tabulation, period (.) and comma (,).

2.4.3. Stopwords removal (filtering)

Stopwords removal (filtering) is removing words that occur frequently but have no meaning [18][6]. Prepositions and conjunctions are also included in stopwords removal. In this study, stopwords removal (filtering) refers to the research proposed by Tala [19].

2.4.4. Stemming

Stemming is useful for returning words to their basic words, it will greatly reduce the kinds of words that need to be checked and compared with stopword tables and keyword tables. This will speed up the word comparison process and reduce the contents of the stopword and keyword tables. The stemming technique for Indonesian began with the development of the Nazief-Adriani algorithm in 1996 [20].

3. Methodology

This research uses the Jaccard similarity method and combines the jaccard similarity with the keyword similarity. Figure 1 shows the steps in this research.

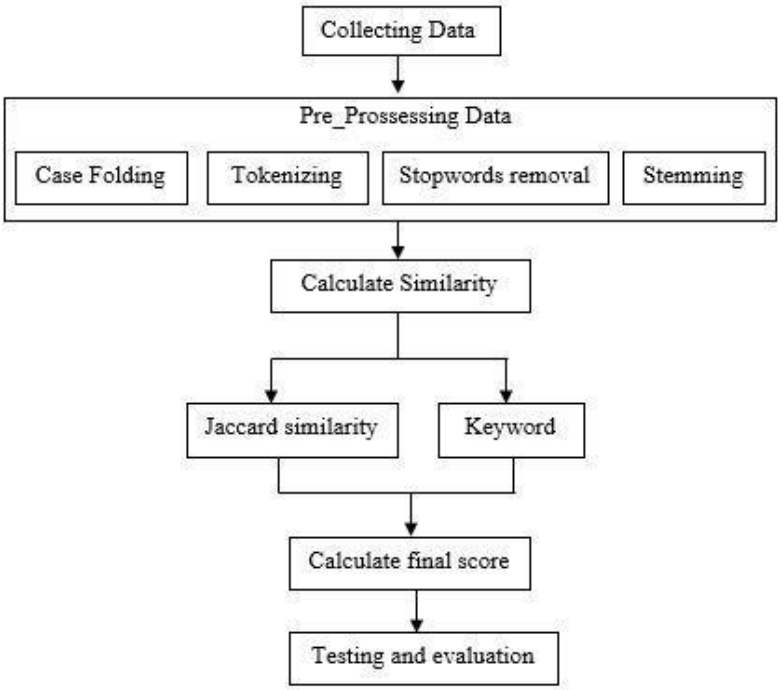

Figure. 1. Research Steps

Data collection is done by choosing a description question in the form of a question that explains the definition or understanding along with an answer key as a reference answer, then collects student answers and the results of the lecturer's assessment. The results of the assessment are used to calculate the accuracy of the description answer scoring system automatically. The student's answers and the lecturer's reference answers did not experience any changes in the writing format, sentence structure, punctuation marks, abbreviations and other elements contained in the student answers.

4. Result and Discussion

4.1. Collecting Data

The research materials used were questions and answers for the Indonesian E-Business course quiz questions at Amikom University, Purwokerto. This data consists of 4 questions with 2 lecturer reference answers. Each question was answered by 31 students so that the total number of student answers was 124. The number of reference answers was determined by the lecturer with the provision that the answers used as references represented the answers to the

questions asked. Students' answers are then scored by two lecturers and the reference score is taken from the average of the two lecturers' scores. Table 1 is an example of a snippet of a dataset of questions, lecturers' answers and student answers.

Table. 1. Dataset Sample using Bahasa Indonesia

Question	Apakah yang dimaksud dengan E-business?	Score
Reference 1	Kegiatan bisnis yang dilakukan secara elektronik dengan menggunakan internet	4
Reference 2	Kegiatan bisnis yang dilakukan secara otomatis dan semi otomatis dengan menggunakan internet	4
Student Answer 1	kegiatan bisnis yang dilakukan secara elektronik atau dengan menggunakan internet	4
Student Answer 2	kegiatan bisnis yang secara elektronik dengan menggunakan internet	4
Student Answer 3	e-Business adalah kegiatan bisnis secara elektronik	0,75

4.2. Preprocessing Data

Before the data is processed, the data first goes through the preprocessing stage. This stage is the stage where the application selects the data to be processed in each document. This preprocessing process includes case folding, tokenizing, Stopwords removal and stemming. At the stemming stage, the stemmer used is the Sastrawi Stemmer which can be downloaded at https://github.com/har07/PySastrawi.). The algorithm used in the Stemmer Sastrawi library is based on the Nazief-Adriani Algorithm, as well as the algorithm from research conducted by Jelita [21][7][20]. In the Stemmer Sastrawi library there is also a feature to do case folding. So that these two processes are carried out simultaneously with the stemming process. Figure 2 shows the stages of data preprocessing.

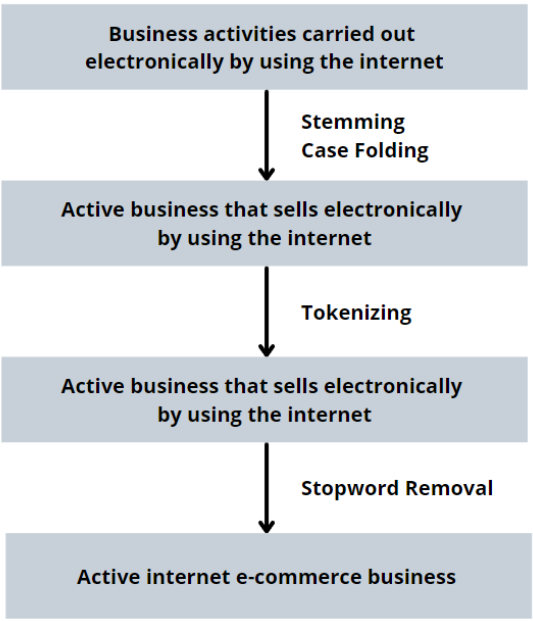

Figure. 2. Data Preprocessing Stage

4.3. Calculate Similarity

After going through the data preprocessing stage, the Jaccard similarity algorithm and keyword similarity were implemented using the Python programming language using Spyder IDE. The calculation of the Jaccard Similarity

algorithm is carried out using equation 1. The following is a sample calculation to find the similarity value using the Jaccard Similarity algorithm and keyword similarity between Lecturer answers (A) and student answers (B).

The similarity value between alternative answer 1 and student answers:

A = active internet electronic business

B = active internet e-commerce business

$$Sim_{Jaccard} = \frac{5}{5} = 1{,}00 \quad (2)$$

$$Sim_{Keyword} = \frac{5}{5} = 1{,}00 \quad (3)$$

The similarity value between alternative answers 2 and student answers:

A = semi-internet automatic business activity

B = active internet e-commerce business

$$Sim_{Jaccard} = \frac{4}{7} = 0{,}57 \quad (4)$$

$$Sim_{Keyword} = \frac{4}{6} = 0{,}67 \quad (5)$$

According to the calculation results above, the Jaccard Similarity Algorithm and similarity keyword give the best results for alternative answer 1 with a similarity value of 1.00 for the student answers above. Table 2 is the calculation of Jaccard similarity using 1 answer reference and 2 alternative answer references in question number 1.

Table. 2. Calculation of Jaccard similarity in Problem Number 1

Student	1 Answer Reference	2 Answer Reference	Best result
1	1,00	0,57	1,00
2	0,80	0,50	0,80
3	0,60	0,33	0,60
4	0,43	0,29	0,43
5	0,80	0,50	0,80
6	0,31	0,43	0,43
7	1,00	0,57	1.00
8	0,80	0,50	0,80
9	0,80	0,57	0,80
10	1,00	0,57	1.00
…	…	…	…
31	0,43	0,29	0,43

Based on Table 2, from 31 student answers, there are 6 results that are worth 1, the rest of the values are 0.20 to 0.80. A value of 1.0 means that the answer obtained is 100% correct, while the next value is varied. Furthermore, the 1.0 value obtained from the reference answer 1 is the student with serial number 1, 7, 10, 13 and 24. Then the 1.0 value obtained from the reference answer 2 is the student with serial number 15. The average value obtained in Table 2 is 0.66. Table 3 is a calculation of Keyword similarity using 1 answer reference and 2 alternative answer references in question number 1.

Table. 3. Calculation of Keyword Similarity in Number 1

Student	1 Answer Reference	2 Answer Reference	Best result

1	1,00	0,67	1,00
2	0,80	0,50	0,80
3	0,60	0,33	0,60
4	0,60	0,33	0,60
5	0,80	0,50	0,80
6	0,80	0,83	0,83
7	0,80	0,50	0,80
8	0,80	0,50	0,80
9	0,80	0,67	0,80
10	1.00	0,67	1.00
...	...	...	...
31	0,80	0,50	0,80

Based on table 3, from 31 student answers, there are 4 results that are worth 1, the rest of the values are 0.20 to 0.83. A value of 1.0 means that the answer obtained is 100% correct, while the next value is varied. Furthermore, the 1.0 value obtained from the reference answer 1 is Student serial number 1, 10, 13 and 24. Then the value is 1.0. In answer 2 reference there is no value 1.0. The average value obtained in Table 3 is 0.70.

4.4. Calculate final Score

The final value is obtained from the average value between the best value for the Jaccard similarity algorithm and the best value for the keyword similarity. Each value is multiplied by the answer score, which is 4. Table 4 is a calculation of the combination of the Jaccard Similarity Algorithm and Keyword.

Table. 4. Calculation of the combination of the Jaccard similarity and Keyword Algorithm

Student	Question 1	Question 2	Question 3	Question 4
1	4,00	2,00	3,43	1,67
2	3,20	2,67	4,00	1,24
3	2,40	1,33	2,29	4,00
4	2,06	4,00	2,86	4,00
5	3,20	3,43	2,86	4,00
6	2,52	4,00	3,43	4,00
7	3,60	3,10	0,00	1,67
8	3,20	2,53	3,43	3,33
9	3,20	4,00	2,29	2,70
10	4,00	2,67	1,71	3,60
...	...	...	...	...
31	2,46	4,00	3,43	1,86

4.5. Testing and Evaluation

After calculating the similarity value, then testing is carried out. The test is carried out by calculating the Pearson correlation value between the similarity value generated by the system and the value given by the lecturer. In addition to the Pearson correlation test, it was also tested using MAE (Means Absolute Error). The comparison of correlation and MAE is shown in Table 5 below:

Table. 5. Comparison of correlation and MAE

Question	Alternative Answer 1		Alternative Answer 2		Jaccard Similarity + Keyword Similarity	
	r	MAE	r	MAE	r	MAE

1	0,69	0,68	0,75	0,62	0,75	0,65
2	0,65	0,80	0,67	0,50	0,75	0,39
3	0,76	0,66	0,85	0,47	0,86	0,46
4	0,67	0,74	0,70	0,60	0,77	0,46
Rata-Rata	**0,69**	**0,72**	**0,74**	**0,55**	**0,78**	**0,49**

The comparison of correlation and MAE in the table above can be summarized in Figure 3.

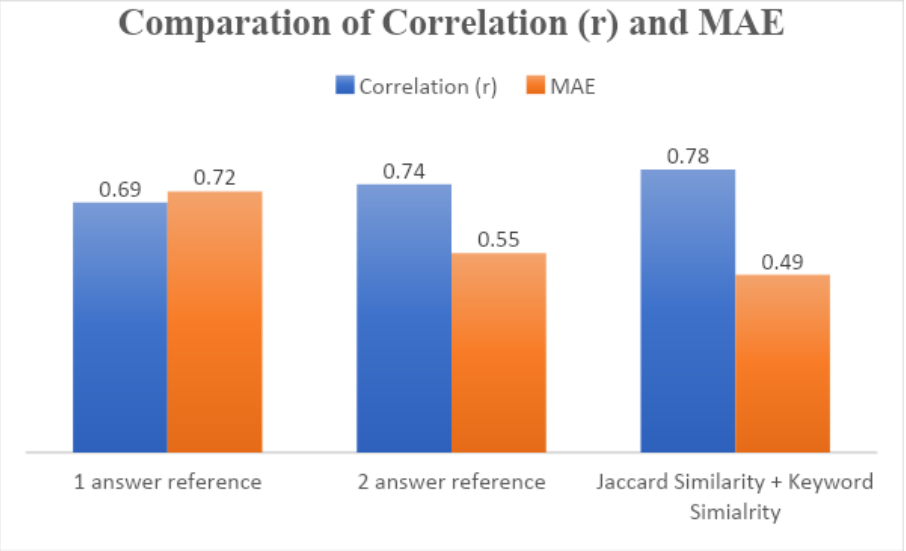

Figure. 3. Comparison of Correlation and MAE

Based on the criteria for the success of the correlation, the value is declared to have a very good correlation value. Furthermore, in the jaccard similarity algorithm, using 1 answer reference produces a correlation of 0.69, which means that it is included in the good category. For the results of the 2 reference answers, it produces a correlation value of 0.74, further after being combined with the similarity keyword it has a correlation value of 0.78. This shows that with 2 references and after a combination of the algorithm has an increase in the correlation value obtained. Based on the criteria for the success of the correlation, the value is declared to have a very good correlation value. Table 5 above also shows that adding alternative answers can reduce the MAE value, as well as after combining it with the MAE value keyword similarity.

5. Conclusion

In this study, the lecturer added an answer reference as an alternative answer. The purpose of adding this answer reference is to overcome variations in student answers. Student answers that are too long can affect similarity, namely the low similarity value when implemented. This research also adds the conventional keyword similarity method. Keyword similarity is a way to reduce words that are not the constituent points of the answer. By using the keyword similarity method, the system only assesses the presence or absence of keywords that are the points of making sentences. Jaccard Similarity Algorithm the average correlation value using 1 answer reference is 0.69 and the MAE average is 0.72. After adding the answer reference as an alternative, it was able to increase the average value of correlation to 0.74 and decrease the average value of MAE to 0.55. Likewise, after being developed by combining the Jaccard Similarity algorithm and keyword matching, the average correlation value increased to 0.78 and the average MAE value decreased to 0.49.

References

[1] Ratna, A.A.P., Budiardjo, B., Hartanto, D., 2010, Simple: Sistem Penilai Esai Otomatis Untuk Menilai Ujian Dalam Bahasa Indonesia', MAKARA, TEKNOLOGI, 11, 5–11.

[2] Hasanah, U., Mutiara, D.A., 2019, Perbandingan Metode Cosine Similarity dan Jaccard Similarity Untuk Penilaian Otomatis Jawaban Pendek, Seminar Nasional Sistem Informasi dan Teknik Informatika, 1255–1263. Available at: https://ejurnal.dipanegara.ac.id/index.php/sensitif/article/view/511

[3] Valenti, S., Neri, F., Cucchiarelli, A., 2003, An Overview of Current Research on Automated Essay Grading, Journal of Information Technology Education: Research, 2, 319–330.

[4] Agarwal, S., 2014, Data mining: Data mining concepts and techniques, Proceedings - 2013 International Conference on Machine Intelligence Research and Advancement, ICMIRA 2013.

[5] Nisbet, R., Miner, G., Yale, K., 2018, Handbook of Statistical Analysis and Data Mining Applications. Elsevier

[6] Wahyuningsih, T., Henderi., Winarno., 2021, Text Mining an Automatic Short Answer Grading (ASAG), Comparison of Three Methods of Cosine Similarity, Jaccard Similarity and Dice's Coefficient', Journal of Applied Data Sciences, 2(2), pp. 45–54. doi: 10.47738/jads.v2i2.31.

[7] Arifin, Z., 2009, Evaluasi Pembelajaran: Prinsip, Teknik, Prosedur, Bandung, Remaja Rosdakarya

[8] Mardapi, D., 2008, Teknik Penyusunan Instrumen Tes dan Nontes Yogyakarta, Mitra Cendikia.

[9] Rusdiana, H., Sumardi, K., Arifiyanto, E.S., 2016, Evaluasi Hasil Belajar Menggunakan Penilaian Autentik Pada Mata Pelajaran Kelistrikan Sistem Refrigerasi, Journal of Mechanical Engineering Education, 1. 274-283

[10] Sudijono, A., 1996, Pengantar Evaluasi Pendidikan, Jakarta, Raja Grafindo Persada.

[11] Arifin, Z., 2012, Menganalisis Kualitas Tes, Evaluasi Pembelajaran, Bandung, Remaja Rosdakarya

[12] Rinartha, K., 2017, Simple Query Suggestion Untuk Pencarian Artikel Menggunakan Jaccard Similarity, Jurnal Ilmiah Rekayasa dan Manajemen Sistem Informasi, 3, 30–34

[13] Fadelillah, M., Subroto, I.M.I., Kurniadi, D., 2017, Sistem Rekomendasi Hasil Pencarian Artikel Menggunakan Metode Jaccard Coefficient, Jurnal Elektro & Informatika, 2, 1–14.

[14] Sugiyamto, S., Surarso, B., Sugiharto, A., 2014, Analisa Performa Metode Cosine dan Jaccard Pada Pengujian Kesamaan Dokumen, Jurnal Masyarakat Informatika, 5, 1–8.

[15] Rahutomo, F., Ayatullah A.H., 2018, Indonesian Dataset Expansion of Microsoft Research Video Description Corpus and Its Similarity Analysis, Kinetik, 3, 319–326.

[16] Dewi, K.E., W. NI., Heryandi, A., 2014, Penilaian Jawaban Essay Menggunakan Semi Discrete Decomposition Pada Metode Latent Semantic Indexing, Seminar Nasional Aplikasi Sains & Teknologi (SNAST), 211–216

[17] Shiri, A., 2004, Introduction to Modern Information Retrieval (2nd edition), Library Review, 53, 462–463.

[18] Patel, B., Shah, D., 2013, Significance of Stop Word Elimination in Meta Search Engine, International Conference on Intelligent Systems and Signal Processing, ISSP 2013, 52–55.

[19] Tala, F. Z., 2003, A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia, M.Sc. Thesis, Appendix D, 39–46.

[20] Tahitoe, A.D., Purwitasari, D., 2010, Implementasi Modifikasi Enhanced Confix Stripping Stemmer Untuk Bahasa Indonesia Dengan Metode Corpus Based Stemming, Jurnal Ilmiah, 1–15

[21] Jelita, A., 2007, Effective Techniques for Indonesian Text Retrieval, Ph.D Thesis, 1–286. Available at: https://researchbank.rmit.edu.au/view/rmit:6312.