

Enhancing service excellence: analyzing natural language question answering with advanced cosine similarity

Riza Arifudin, Subhan, Yahya Nur Ifriza

Department of Computer Science, Universitas Negeri Semarang, Semarang, Indonesia

Article Info

Article history:

Received Jan 13, 2023

Revised Nov 3, 2023

Accepted Nov 15, 2023

Keywords:

Cosine similarity
Natural language question answering system
Performance
Statistical scoring

ABSTRACT

Information related to student services in higher education must be produced and disseminated in various forms. Covid-19 pandemic, student services with a remote model related to this question and answer become very important. To carry out this automation process, the advanced cosine similarity method is used to check the similarity of the questions to the database and statistics to calculate the similarity value of each word. The proposed paper proceeds with three phases. The first stage to solve this problem is the data processed in question; the professional next step is word insertion. It converts alphanumeric words to vector format. Each word is a vector that represents a point in space with a certain dimension. The recommended advanced cosine similarity data still must be analyzed into a statistical approach. We will measure accuracy to get results so that optimal results and answers are obtained, research procedures are carried out based on literature study, initial data collection and observation, system development, system testing, system analysis, and system evaluation. This research implemented in universities with student chat automation applications providing an accuracy 83.90% given by natural language question answering system (NLQAS) so that it can improve excellent service in universities.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Riza Arifudin
Department of Computer Science, Universitas Negeri Semarang
Sekaran, Gunungpati, Semarang, Indonesia
Email: rizaarifudin@mail.unnes.ac.id

1. INTRODUCTION

Education is critical to the development and improvement of human resources. The quality of services provided to users (students and the community) is one of the success determinants of educational institutions [1]. Improving the quality of higher education is a universal obligation that must be carried out by all higher education providers, including government and community-based higher education. As the primary stakeholders in higher education, students should be allowed to acquire what they desire [2], [3]. To ensure that students receive what they anticipate, the institution must be able to align student expectations with the organization's vision, purpose, and goals. Academic services will be carried out to prioritize quality factors, suitable facilities, and competent administration, resulting in a synergy of student expectations and campus interests [3]. Higher education institutions are service-oriented organizations. The quality of higher education institutions is gradually becoming a stakeholder demand in the higher education service industry [4], [5]. The quality that higher education institutions must give is the provision of services that can bring satisfaction, particularly to students [6], [7]. Student services are facing unique problems because of the Covid-19 epidemic. Students must be able to converse effectively online [8], [9]. The significance of service quality in higher education has steadily been recognized, and the function of service quality in higher education has

gained growing emphasis during the previous two decades [9]. This university is involved with a variety of stakeholders. A university is a type of higher education institution that provides academic education in various fields of science and technology [10]. The university is a higher education institution with the broadest field of knowledge. In other words, almost all kinds of knowledge exist in universities. Student status is status at a high intellectual level, young age with strong ideals [11]. In this case, the role and function of students is very much needed for the progress of a nation and state [12]–[14].

In the current era of the digital revolution, information on student services must be developed and delivered in various methods. Remote services are likely during this Covid-19 epidemic [15]–[17]. One of these long-distance services is handling queries and replies for university services [17]. In our digital age, students may ask and answer questions. This necessitates the automation of student service-related queries and responses. This study's questions and responses were delivered via email and others via email [18]. The query-and-solution technique is the transport of instructions using the trainer to ask questions and college students answering [19], [20]. There are weaknesses and strengths in the query-and-solution technique, so a trainer has to be aware of the suitability of the problem depending on the technique to be used. Numerous matters should be considered in using the query-and-solution technique [21]. First, the form of the query; second, the approach of asking questions; third, taking note of the situations for the usage of the query-and-solution technique so that the proper steps may be formulated; fourth, taking note of the standards of the usage of the query and solution technique, such as the standards of harmony, integration, freedom, and individuality [22]. In addition, the query-and-solution technique also can be blended with different methods, consisting of the lecture technique, assignment, and discussion [23].

Natural language question answering system (NLQAS) is a computer system that can automatically answer questions using the natural language that people usually use. The system's responses include data from a database source. This question and response system is a subset of information seeking. A question-and-answer system is a system that tries to locate the pertinent information for a query posed in natural language given a collection of documents. The availability of this automated question and answer system can improve the quality of student services in the form of automatic live chat. Students are often confused about existing policies and provisions, academic, student affairs, and cooperation. It is necessary to have a system that automatically answers quickly and accurately to provide the service process excellently [24]. A way is required to carry out this automation to work smoothly. The advanced cosine similarity approach is reasonably practical for determining the similarity with statistical approach. It is predicted that this technology would recognize the answers that pupils seek when they ask queries.

2. METHOD

Since 2021, we have been analyzing the implementation of NLQAS and prototyping in the computer science major. The method used in implementing NLQAS is the cosine similarity technique and the statistical assessment method. Most of the students will ask their friends if they have difficulties providing services at the university, even though the friends who are asked also do not know the answers to what is being asked. So, we need a system that connects students with academic officers on campus so that answers to problems encountered by students get the right solution through NLQAS [25], [26]. The automatic question answering system framework uses cosine similarity and a statistical approach. Figure 1 shows the automatic question answering system architecture, including the question processing module strategies. Student questions, document processing, message-to-response extraction. The first step in NLQAS architecture is to understand the student's question asked by the user rather than being given the chance to enter the system to find the answer. It can be done using cosine similarity techniques, such as the similarity between two documents [27]. The three most relevant documents using cosine similarity were obtained based on the evaluation value. After ingesting the top 3 documents, the next level is the answer part search module. Each high-level document is divided into different sections, and each section is considered a document. Or the advanced evaluation approach is applied again to some sections of the user's question to extract the exact answer [28].

Data preprocessing is a sequence of parts of the process practiced preparing a data set for analysis and modeling. So, this stage is considered as an important step in the data mining process. In this study, data preprocessing included data cleaning, data normalization, and data recovery, during data cleaning, missing values, inconsistencies, and noise (e.g. incorrect data entry) is removed. We used a student and learning document dataset for documents obtained from universities, containing 2,875 response data from service operations in integrated service units with 21 attributes count. Then, after preprocessing the data, the missing values are filled in through interpolation mode and some effect or no effect properties are removed so we get three attributes question_body, professionals_stakeholders and answers_body displaying the attributes used in the test.

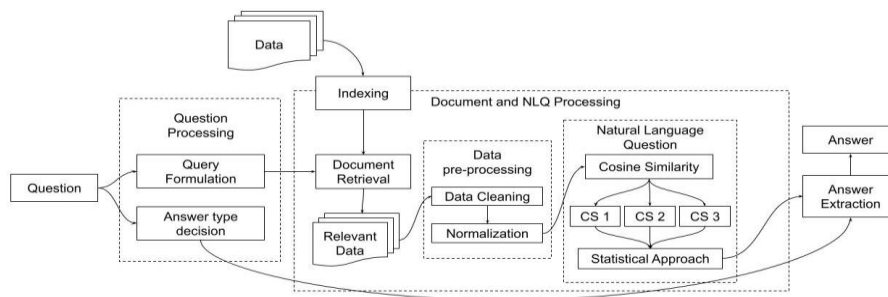


Figure 1. NLQAS architecture

2.1. Cosine similarity

Cosine similarity is measured using cosine similarity, regardless of size. It computes the sine and cosine angles between two vectors projected into three dimensions. Even if two comparable papers are separated by the Euclidean distance (due to document size), cosine similarity is functional. The greater the cosine similarity, the smaller the angle. When applied to multidimensional space, if each dimension corresponds to a word in the document, cosine similarity identifies the document's orientation (angle) rather than its size. If you want to make an order of magnitude, calculate Euclidean distance instead. Even if two similar documents occur for size, even if two similar documents occur due to the size of the size (like the word "cricket," and ten times in other things, they may still have. There is one slight angle between them. The angle decreases and the similarity is high.

To calculate cosine similarity, we need to count the words in each document. We can use the count vector or scikit learn Tfidf vectorizer to calculate this. The problem is as sparse_matrix. We can then convert to Pandas DataFrame and display mono text frequencies in tabular format. We can even use TfidfVectorizer () instead of CountVectorizer (). This is because it contains down score words that frequently appear in the document. Then use cosine_similarity () to get the final output. You can use the document term matrix as a Panda DataFrame and the sparse matrix as input [29]. The similarity of two vectors in the inner product space is measured by cosine similarity. It is calculated by taking the sine and cosine of the angle between the two vectors and determining if they point in the same direction. It is commonly used in text analysis to determine the similarity of documents. A document can be represented by millions of characteristics, each of which records the frequency of a certain word (or phrase) or phrase in the text. As a result, each document is represented by a term frequency vector [30]. For example, in Table 1, document 1 contains the word team five times, while education appears three times. The word ult does not exist throughout the document, as indicated by the count 0. Such data can be very asymmetric.

Table 1. Document vector or term-frequency vector

Document	Education	ult	UKT	Academic	Student
Document1	520	211	310	236	267
Document2	310	534	220	179	563
Document3	225	710	132	245	328
Document4	612	130	341	329	621

Typically, the word frequency vector is rather lengthy (i.e., you have many 0 values). Information calls, text document grouping, biological taxim, and functional genetic assignment are examples of applications that employ such structures. Traditional removal methods discussed in this chapter are ineffective for such economic numeric data. Two-terminal frequency vectors, for example, can be concatenated, implying that the matching document does not share many words, but this is not similar. It is vital to concentrate on words with two documents and the events associated with such terms. In other words, you need a numeric data measure that ignores the zero match. Cosine similarity is a measure of similarity that can be used to compare documents or, for example, rank documents against a particular vector of search terms [31]. For comparison, let x and y be two vectors. Using the cosine measure as a similarity function in (1):

$$\sin(x, y) = \frac{x \cdot y}{||x|| ||y||} \tag{1}$$

where $\|x\|$ is the Euclidean norm of the vector $x = (x_1, x_2, \dots, x_n)$ defined as $\sqrt{x_1^2 + x_2^2 + \dots + x_n^2}$. It is the vector's length in concept. Likewise, $\|y\|$ represents the Euclidean vector y . The cosine of the angle between vectors x and y is computed using the measure. A cosine value of 0 indicates that the two vectors are 90 degrees apart (orthogonal) and do not match. The lower the angle and the higher the match between vectors, the closer the cosine value to 1. Because the cosine similarity measure does not satisfy all of the requirements defined in value obtained, it is referred to as a non-metric measure [32]. In addition, this approach is also used in data mining to quantify cohesive forces between clusters. It can be simple, especially for sparsely populated vectors, as you only must consider non-zero dimensions. Cosine similarity is also known as Urchin similarity and Tucker match factor. Otsuka Chia similarity on (2) is the cosine similarity applied to binary data, and the tucker match factor is another name for cosine similarity. Cosine similarity algorithm describes the semantic similarity of short texts. The cosine of two non-zero vectors may be calculated using the Euclidean dot product formula, the cosine similarity, $\cos(\theta)$, between two vectors of characteristics, A and B , is expressed as a dot product with a magnitude as in (2):

$$\text{cosine similarity} = Sc(A, B) = \cos \theta = \frac{A \cdot B}{\|A\| \|B\|} = \frac{\sum_{i=1}^n A_i \cdot B_i}{\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}} \quad (2)$$

where A_i and B_i are components of vectors A and B , respectively.

2.2. Statistical method

Statistical methods support classification in four ways. When developing a probabilistic model of data and classes to identify possible classifications for a particular dataset. Development of validation tests for specific classes generated by the classification scheme. When evaluating the efficacy of various categorization schemes. Use a probabilistic search method to broaden your search for the best categorization. Standard hierarchy and partitioning methods have been statistically examined, and several improvements to these algorithms have been proposed utilizing density estimates and mixed models. It demonstrates how to determine the multimodality of high-dimensional data [33]. Statistics are the fields of science that handle collection, organization, data analysis of data analysis, and key keys from samples. This requires selecting the appropriate design of the study, the appropriate selection of tests, and appropriate statistical tests. Appropriate knowledge of statistics is required to develop epidemiological research or clinical trials properly. Inappropriate statistical methods may have the disadvantage of leading to unethical implementation. Variables are characteristics that vary from a single member of a population to another individual. Variables such as height and weight are measured on specific scales and convey quantitative information and are called quantitative variables. Gender and eye color provide qualitative information and are called qualitative variables [34]. The spread expresses the degree to which the observed values gather around the center position to the extremum due to the central tendency and spread. Central tendency measurements are mean, median, and mode. The average (or arithmetic mean) is the sum of all reviews divided by the number of reviews. The average can be strongly influenced by extreme variables [35], [36]. For example, the average length of stay in the intensive care unit (ICU) for organophosphate poisoning patients can be affected by one patient staying in the ICU for about five months due to sepsis. Extremes are called outliers. The average formula as in (3):

$$\bar{x} = \frac{\sum x}{n} \quad (3)$$

where x is the number of observations and n is the number of observations. In ranking data, the median is the center of the distribution (with half of the variables in the sample above and half below the median value), whereas the mode is the most often occurring variable in the distribution. The spread or variability is defined by the range [37], [38]. It is defined by the variables' minimum and maximum values. We can learn more about the pattern of dispersion of the variables if we rate the data and classify the observations into percentiles [38], [39]. In percentiles, we rank the observations into 100 equal parts. You can then describe the 25%, 50%, 75%, or another percentile amount. The median is the 50th percentile [40]. The interquartile range corresponds to the median 50% of the observations near the median (25th to 75th percentiles) [41]. Variance is a measure of how wide the distribution is. This shows how close each observation group is to the average [42].

2.3. Accuracy

Accuracy is a metric used to evaluate a classification model. Informally, accuracy is the model's percentage correctly predicted [43]. Formally, accuracy has the following definition. For binary classification, accuracy can also be calculated in terms of positives and negatives as in (4):

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

where TP denotes true positives, TN denotes true negatives, FP denotes false positives, and FN denotes false negatives. TP and FN are shown in green since they were accurately anticipated observations. Reduce the number of FP and FP such that they are highlighted in red. These words are a little perplexing. Let's go down each concept and completely comprehend it. TP values are those that were accurately anticipated to be positive. The actual grade value is yes, as is the anticipated grade value. For example, if the actual class value indicates that this passenger survived, and the anticipated class value also suggests that this passenger survived. TN are negative numbers that were accurately anticipated. There are FN and FP. For instance, if the actual class value reveals that this passenger has survived, and the forecast class suggests that the passenger will die [44].

The simplest basic metric of performance is accuracy, which is just the ratio of properly predicted observations to total number of observations [45]. As a result, you must consider additional criteria while evaluating the performance of your model. We have 0.803 for our model. This model is around 80% accurate [46]. Accuracy can be determined by one measurement or experiment, but to determine accuracy requires measurement to assess precision [47]–[49]. Accuracy measurements can be accurate but not necessarily exact [50], [51]. Accurate value in accuracy is affected by the degree of conformity. Meanwhile, for precision measurements can be done once, but the results are not necessarily exact [52], [53]. Because it is necessary to determine server measurements. Precision may suffer random errors, so that exact values may or may not be accurate and are affected by the degree of reproducibility [54]. In making measurements, accuracy and precision are very important [55]. Because precision and accuracy will affect the appropriate size and produce the right size results [56]. In addition, a measuring instrument must have high accuracy and precision to reduce the occurrence of errors in measurement [57].

3. RESULTS AND DISCUSSION

This study resulted in a live chat automation solution for student services using NLQAS model. The author creates an application using an information system framework in this application. This framework is also built into an application that handles student services. Researchers design an information system framework that will be established when developing the information system framework. Figure 1 depicts the framework of the system to be created. These stages will be described as follows.

3.1. Cosine similarity

This study developed an NLQAS application utilizing the cosine similarity approach and data from stakeholders. The author creates an application using an information system framework. The processed data is in questions, professionals, and answers. Furthermore, the data is preprocessed to decide which variables will be utilized. The data is translated from word to vector. The data is processed using a cosine similarity matrix to discover which suggestions are comparable to the questions asked. Based on preprocessing, we can determine the variables we choose, as shown in Table 2.

The next step is word embedding. This converts alphanumeric words to vector format. Each word is a vector that represents a point in space with a particular dimension. Words that share a particular characteristic. This process, recommended answers are obtained from a professional that students can use to determine what steps should be taken to resolve the problem. Problems that usually have to be resolved by meeting with administrative officers and can find solutions quickly. From the question how to pay for single tuition fee/*uang kuliah tunggal* (UKT) and institutional development contribution/*sumbangan pengembangan institusi* (SPI)? with cosine similarity the process gives results three answers, more details can be seen in Table 3.

Table 2. Variables selected by preprocessing

No	Feature name	Description	type
1	questions_body	Expression of someone's curiosity about information that is contained in a question sentence.	object
2	professionals_stakeholders	Person who offers services or services in accordance with protocols and regulations in the field he is in	object
3	answers_body	Response or reply; something said or done in reaction to a statement or question.	object

Table 3. Agency of quality assurance recommendation answer

Question	Agency of quality assurance answer	Cosine similarity
How to pay for UKT and SPI	Payment steps...	0.83
	Terms and Conditions...	0.76
	Steps that need to be taken...	0.75

When using word embedding, they are in the same context or have the same meaning, not separated by this space. In this scenario, add a query statement such as "how to pay for UKT and SPI?". For the word2vec calculation, the simulation is shown in Figure 2.

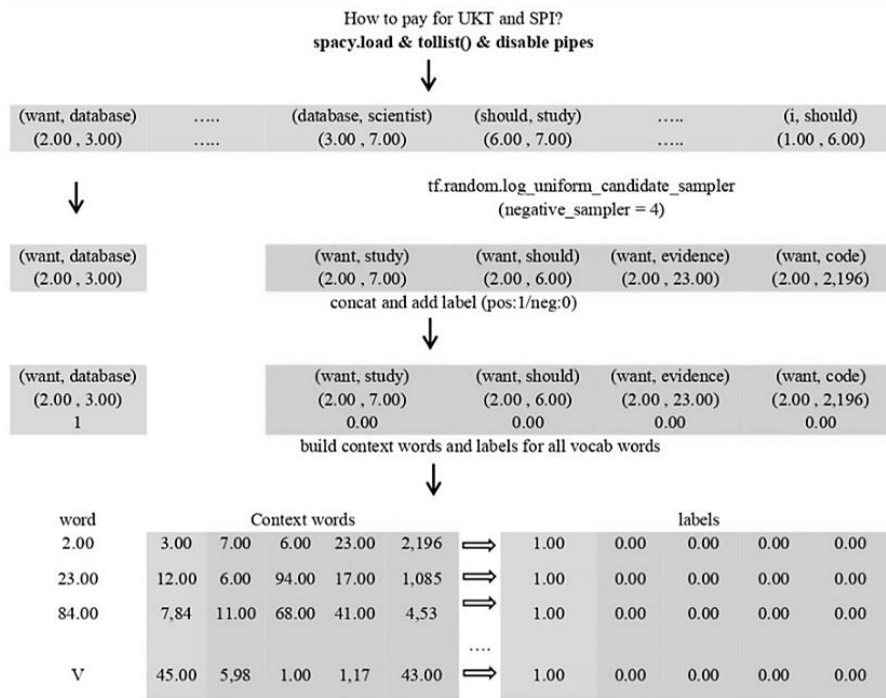


Figure 2. word2vec simulation

3.2. Statistical approach

The recommendation data on cosine similarity still has to be analyzed into a statistical approach. This is so that students, when given an answer from NLQAS, are no longer confused with one definite answer. From the testing experiments three recommendations for cosine similarity were given in the form of CS1, CS2, and CS3. The statistical approach process can be shown in Figure 3.

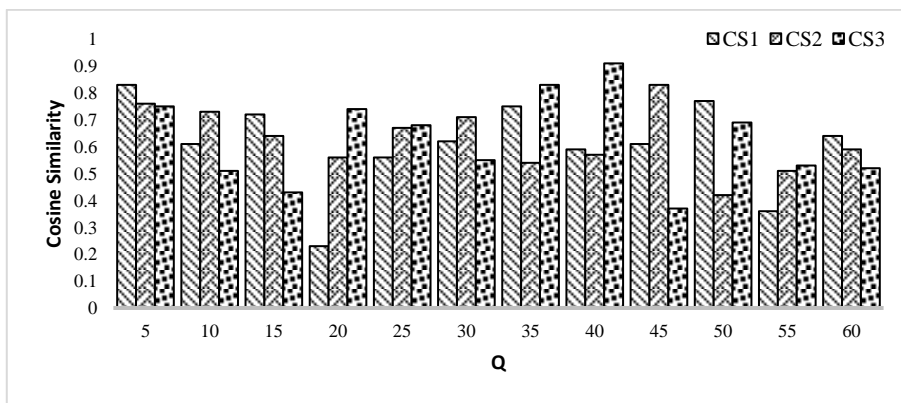


Figure 3. The advanced cosine similarity recommendation

3.3. Accuracy

Based on the trials we did on the application, it is known from the test table data that three out of 12 cases failed, performance measures may be done using the accuracy and error rate formulation, we will

measure accuracy to get results that are as close as possible to the actual results so that optimal results are obtained. The calculation results. In the study before using the advanced cosine similarity and after completing computations using a dataset data, the accuracy was found 77.60% and then improved to be 83.90%. These results, when compared with previous studies regarding the evaluation of student responses, were achieved using a multi-criteria decision-making approach. It uses a set of model answers drawn from various textbooks and subject experts to evaluate answers. Various measures were used to assess answers by comparing them with this model set [14]. The results of the in-depth system reveal that the automatic assessment process can reduce manual human effort which has not yielded measurable results, so that the novelty in this study by combining the two methods provides innovation in providing novelty solutions within the framework of natural language questions.

4. CONCLUSION

In research on the implementation of the cosine similarity and statistical approach method using the higher education dataset to predict correct answer from NLQAS, it can be concluded as follows: the cosine similarity and statistical approach method can be used to predict correct answer from NLQAS by considering the factors that affect the question. The results of the accuracy of calculations using advanced cosine similarity method to predict correct answers from NLQAS are 83.90%. We can easily extend this same technique to other text-based case studies for a variety of applications, such as recommendations for answers in a call center, recommendations for donors who have donated for a social cause in charitable organizations, and recommendations for donors who have donated for a social cause in for-profit businesses.

ACKNOWLEDGEMENTS

This research is supported by the budget implementation list (DIPA) of Universitas Negeri Semarang Number: SP DIPA-023.17.2.677507/2021, November 23, 2020 under letter of assignment for the implementation of basic research (University) 2021 UNNES DIPA funds.

REFERENCES




- [1] N. Othman, R. Faiz, and K. Smaïli, "Learning English and Arabic question similarity with Siamese Neural Networks in community question answering services," *Data & Knowledge Engineering*, vol. 138, 2022, doi: 10.1016/j.datak.2021.101962.
- [2] J. Yin and S. Sun, "Incomplete multi-view clustering with cosine similarity," *Pattern Recognition*, vol. 123, 2022, doi: 10.1016/j.patcog.2021.108371.
- [3] M. M. A. Mahfouz, "A protection scheme for multi-distributed smart microgrid based on auto-cosine similarity of feeders current patterns," *Electric Power Systems Research*, vol. 186, no. 18, pp. 1–9, 2020, doi: 10.1016/j.epsr.2020.106405.
- [4] Y. N. Ifriza, C. E. Edi, and J. E. Suseno, "Expert system irrigation management of agricultural reservoir system using analytical hierarchy process (AHP) and forward chaining method," in *Proceeding of ICMSE*, 2017, vol. 4, no. 1, pp. 74–83.
- [5] M. Hanifi, H. Chibane, R. Houssin, and D. Cavallucci, "Problem formulation in inventive design using Doc2vec and Cosine Similarity as Artificial Intelligence methods and Scientific Papers," *Engineering Applications of Artificial Intelligence*, vol. 109, pp. 1–38, 2022, doi: 10.1016/j.engappai.2022.104661.
- [6] N. Arunachalam and A. Amuthan, "Improved Cosine Similarity-based Artificial Bee Colony Optimization scheme for reactive and dynamic service composition," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 2, pp. 270–281, 2022, doi: 10.1016/j.jksuci.2018.10.003.
- [7] J. Chen, Z. Guo, and J. Hu, "Ring-Regularized Cosine Similarity Learning for Fine-Grained Face Verification," *Pattern Recognition Letters*, vol. 148, pp. 68–74, 2021, doi: 10.1016/j.patrec.2021.04.029.
- [8] D. Fonseca, S. Villagrasa, N. Martí, E. Redondo, and A. Sánchez, "Visualization Methods in Architecture Education Using 3D Virtual Models and Augmented Reality in Mobile and Social Networks," *Procedia - Social and Behavioral Sciences*, vol. 93, pp. 1337–1343, 2013, doi: 10.1016/j.sbspro.2013.10.040.
- [9] H. Kalhori, M. M. Alamdari, and L. Ye, "Automated algorithm for impact force identification using cosine similarity searching," *Measurement: Journal of the International Measurement Confederation*, vol. 122, pp. 648–657, 2018, doi: 10.1016/j.measurement.2018.01.016.
- [10] W. Hu, L. Wu, M. Jian, Y. Chen, and H. Yu, "Cosine metric supervised deep hashing with balanced similarity," *Neurocomputing*, vol. 448, pp. 94–105, 2021, doi: 10.1016/j.neucom.2021.03.093.
- [11] M. Abdel-Basset, M. Mohamed, M. Elhoseny, L. H. Son, F. Chiclana, and A. E.-N. H. Zaied, "Cosine similarity measures of bipolar neutrosophic set for diagnosis of bipolar disorder diseases," *Artificial Intelligence in Medicine*, vol. 101, pp. 1–31, 2019, doi: 10.1016/j.artmed.2019.101735.
- [12] N. Grieb, T. Oltrup, T. Bende, and M. A. Leitritz, "The Cosine Similarity Technique: A new method for smart EXCIMER laser control," *Zeitschrift für Medizinische Physik*, vol. 30, no. 4, pp. 253–258, 2020, doi: 10.1016/j.zemedi.2020.02.006.
- [13] B. Il Kwak, M. L. Han, and H. K. Kim, "Cosine similarity based anomaly detection methodology for the CAN bus," *Expert Systems with Applications*, vol. 166, 2021, doi: 10.1016/j.eswa.2020.114066.
- [14] B. Das, M. Majumder, A. A. Sekh, and S. Phadikar, "Automatic question generation and answer assessment for subjective examination," *Cognitive Systems Research*, vol. 72, pp. 14–22, 2022, doi: 10.1016/j.cogsys.2021.11.002.
- [15] J. Y. Dong, Y. Chen, and S. P. Wan, "A cosine similarity based QUALIFLEX approach with hesitant fuzzy linguistic term sets for financial performance evaluation," *Applied Soft Computing Journal*, vol. 69, pp. 316–329, 2018, doi: 10.1016/j.asoc.2018.04.053.
- [16] A. L. M. O. C. Torres, "Understanding and intervening in E-learning in higher education institution," *Procedia - Social and*

- Behavioral Sciences*, vol. 15, pp. 756–760, 2011, doi: 10.1016/j.sbspro.2011.03.178.
- [17] V. Bleotu, “Comparative Analysis of Romanian Competitiveness Evolution,” *Procedia - Social and Behavioral Sciences*, vol. 46, pp. 5382–5386, 2012, doi: 10.1016/j.sbspro.2012.06.443.
- [18] M. Ana-Andreea, N. M. Liviu, and M. C. Alina, “Factors of Influence in the Choice of a Higher Education Specialization in Romania,” *Procedia - Social and Behavioral Sciences*, vol. 84, pp. 1041–1044, 2013, doi: 10.1016/j.sbspro.2013.06.695.
- [19] I. B. Ardashkin, “Philosophy of Education as a Social Development Factor: World Trends and Prospects for Russia,” *Procedia - Social and Behavioral Sciences*, vol. 166, pp. 277–286, 2015, doi: 10.1016/j.sbspro.2014.12.524.
- [20] S. Ozkazanc and U. D. Yuksel, “Evaluation of Disaster Awareness and Sensitivity Level of Higher Education Students,” *Procedia - Social and Behavioral Sciences*, vol. 197, pp. 745–753, 2015, doi: 10.1016/j.sbspro.2015.07.168.
- [21] E. Munastiwi, “The Management Model of Vocational Education Quality Assurance Using ‘Holistic Skills Education (Holsked),’” *Procedia - Social and Behavioral Sciences*, vol. 204, pp. 218–230, 2015, doi: 10.1016/j.sbspro.2015.08.144.
- [22] Z. H. U. Aiqun, “An IT capability approach to informatization construction of higher education institutions,” *Procedia Computer Science*, vol. 131, pp. 683–690, 2018, doi: 10.1016/j.procs.2018.04.312.
- [23] N. Othman, R. Faiz, and K. Smaïli, “Enhancing question retrieval in community question answering using word embeddings,” *Procedia Computer Science*, vol. 159, pp. 485–494, 2019, doi: 10.1016/j.procs.2019.09.203.
- [24] N. Süzen, A. N. Gorban, J. Levesley, and E. M. Mirkes, “Automatic short answer grading and feedback using text mining methods,” *Procedia Computer Science*, vol. 169, pp. 726–743, 2020, doi: 10.1016/j.procs.2020.02.171.
- [25] Y. Long, W. Zhao, and L. Chen, “A multi-objective tool selection method using FAHP and cosine similarity,” *Procedia CIRP*, vol. 104, pp. 1843–1848, 2021, doi: 10.1016/j.procir.2021.11.311.
- [26] V. B. Marcus, N. A. Atan, S. M. Salleh, L. M. Tahir, and S. M. Yusof, “Exploring Student Emotional Engagement in Extreme E-service Learning,” *International Journal of Emerging Technologies in Learning*, vol. 16, no. 23, pp. 43–55, 2021, doi: 10.3991/ijet.v16i23.27427.
- [27] H. Lajane *et al.*, “A Scenario of the Formative E-assessment Based on the ARCS Model: What Is the Impact on Student Motivation in Educational Context?,” *International Journal of Emerging Technologies in Learning*, vol. 16, no. 24, pp. 135–148, 2021, doi: 10.3991/ijet.v16i24.24121.
- [28] T. N. Manjunath, D. Yogish, S. Mahalakshmi, and H. K. Yogish, “Smart question answering system using vectorization approach and statistical scoring method,” *Materials Today: Proceedings*, vol. 80, pp. 3719–3725, 2023, doi: 10.1016/j.matpr.2021.07.369.
- [29] J. H. Lau and T. Baldwin, “Practical Insights into Document Embedding Generation,” in *Proceedings of the 1st Workshop on Representation Learning for NLP*, 2014, pp. 78–86.
- [30] L. Wendlandt, J. K. Kummerfeld, and R. Mihalcea, “Factors Influencing the Surprising Instability of Word Embeddings,” in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2018, vol. 1, pp. 2092–2102, doi: 10.18653/v1/N18-1190.
- [31] D. A. Koutsomitropoulos, A. D. Andriopoulos, and S. D. Likothanassis, “Semantic classification and indexing of open educational resources with word embeddings and ontologies,” *Cybernetics and Information Technologies*, vol. 20, no. 5, pp. 95–116, 2020, doi: 10.2478/cait-2020-0043.
- [32] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, “Item-based collaborative filtering recommendation algorithms,” in *Proceedings of the 10th international conference on World Wide Web*, 2001, pp. 285–295, doi: 10.1145/371920.372071.
- [33] L. Zheng, V. Noroozi, and P. S. Yu, “Joint Deep Modeling of Users and Items Using Reviews for Recommendation,” in *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, 2017, pp. 425–434, doi: 10.1145/3018661.3018665.
- [34] D. J. Steele, “Discussion of application of the classifications and group index in estimating desirable subbase and total pavement thicknesses,” in *Highway Research Board Proceedings*, 1946, vol. 25, pp. 388–392.
- [35] G. d. S. P. Moreira, F. Ferreira, and A. M. da Cunha, “News Session-Based Recommendations using Deep Neural Networks,” in *Proceedings of the 3rd Workshop on Deep Learning for Recommender Systems*, 2018, pp. 15–23, doi: 10.1145/3270323.3270328.
- [36] O. Barkan and N. Koenigstein, “ITEM2VEC: Neural item embedding for collaborative filtering,” in *2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*, 2016, pp. 1–6, doi: 10.1109/MLSP.2016.7738886.
- [37] W. B. Zulfikar, M. Irfan, M. Ghufro, Jumadi, and E. Firmansyah, “Marketplace affiliates potential analysis using cosine similarity and vision-based page segmentation,” *Bulletin of Electrical Engineering and Informatics*, vol. 9, no. 6, pp. 2492–2498, 2020, doi: 10.11591/eei.v9i6.2018.
- [38] N. A. Hisham, S. A. Z. S. Salim, A. Hagishima, F. Yakub, and H. F. S. Saipol, “Statistical analysis of air-conditioning and total load diversity in typical residential buildings,” *Bulletin of Electrical Engineering and Informatics*, vol. 10, no. 1, pp. 1–9, 2020, doi: 10.11591/eei.v10i1.2299.
- [39] O. Levy, Y. Goldberg, and I. Dagan, “Improving Distributional Similarity with Lessons Learned from Word Embeddings,” *Transactions of the Association for Computational Linguistics*, vol. 3, pp. 211–225, 2015, doi: 10.1162/tacl_a_00134.
- [40] J. L. Schmidt *et al.*, “Estimating the relative frequency of leukodystrophies and recommendations for carrier screening in the era of next-generation sequencing,” *American Journal of Medical Genetics, Part A*, vol. 182, no. 8, pp. 1906–1912, 2020, doi: 10.1002/ajmg.a.61641.
- [41] A. Karnik, S. Goswami, and R. Guha, “Detecting Obfuscated Viruses Using Cosine Similarity Analysis,” in *First Asia International Conference on Modelling & Simulation (AMS’07)*, 2007, pp. 165–170, doi: 10.1109/AMS.2007.31.
- [42] A. R. Lahitani, A. E. Permasari, and N. A. Setiawan, “Cosine similarity to determine similarity measure: Study case in online essay assessment,” in *2016 4th International Conference on Cyber and IT Service Management*, 2016, pp. 1–6, doi: 10.1109/CITSM.2016.7577578.
- [43] N. Dehak, R. Dehak, J. Glass, D. Reynolds, and P. Kenny, “Cosine similarity scoring without score normalization techniques,” *Odyssey 2010: Speaker and Language Recognition Workshop*, pp. 71–75, 2010.
- [44] D. Gunawan, C. A. Sembiring, and M. A. Budiman, “The Implementation of Cosine Similarity to Calculate Text Relevance between Two Documents,” *Journal of Physics: Conference Series*, vol. 978, no. 1, pp. 1–6, 2018, doi: 10.1088/1742-6596/978/1/012120.
- [45] J. Ye, “Improved cosine similarity measures of simplified neutrosophic sets for medical diagnoses,” *Artificial Intelligence in Medicine*, vol. 63, no. 3, pp. 171–179, 2015, doi: 10.1016/j.artmed.2014.12.007.
- [46] S. Sohangir and D. Wang, “Improved sqrt-cosine similarity measurement,” *Journal of Big Data*, vol. 4, no. 1, pp. 1–13, 2017, doi: 10.1186/s40537-017-0083-6.
- [47] W. Cardoso and R. Di Felice, “Prediction of silicon content in the hot metal using Bayesian networks and probabilistic reasoning,” *International Journal of Advances in Intelligent Informatics*, vol. 7, no. 3, pp. 268–281, 2021, doi: 10.26555/ijain.v7i3.771.




- [48] N. A. Othman, C. F. M. Foozy, A. Mustapha, S. A. Mostafa, S. Palaniappan, and S. A. Kashinath, "A data mining approach for classification of traffic violations types," *International Journal of Advances in Intelligent Informatics*, vol. 7, no. 3, pp. 282–291, 2021, doi: 10.26555/ijain.v7i3.708.
- [49] R. Hidayat, I. T. R. Yanto, A. A. Ramli, and M. F. M. Fudzee, "Similarity measure fuzzy soft set for phishing detection," *International Journal of Advances in Intelligent Informatics*, vol. 7, no. 1, pp. 101–111, 2021, doi: 10.26555/ijain.v7i1.605.
- [50] M. I. Prasetiyowati, N. U. Maulidevi, and K. Surendro, "Feature selection to increase the random forest method performance on high dimensional data," *International Journal of Advances in Intelligent Informatics*, vol. 6, no. 3, pp. 303–312, 2020, doi: 10.26555/ijain.v6i3.471.
- [51] Z. Su, X. Zheng, J. Ai, Y. Shen, and X. Zhang, "Link prediction in recommender systems based on vector similarity," *Physica A: Statistical Mechanics and its Applications*, vol. 560, pp. 1–12, 2020, doi: 10.1016/j.physa.2020.125154.
- [52] M. Luo and Y. Zhang, "A new similarity measure between picture fuzzy sets and its application," *Engineering Applications of Artificial Intelligence*, vol. 96, pp. 1–9, 2020, doi: 10.1016/j.engappai.2020.103956.
- [53] M. Fang, A. Jandigulov, Z. Snezhko, L. Volkov, and O. Dudnik, "New Technologies in Educational Solutions in the Field of STEM: The Use of Online Communication Services to Manage Teamwork in Project-Based Learning Activities," *International Journal of Emerging Technologies in Learning*, vol. 16, no. 24, pp. 4–22, 2021, doi: 10.3991/ijet.v16i24.25227.
- [54] S. A. S. Shishavan, F. K. Gündoğdu, E. Farrokhzadeh, Y. Donyatalab, and C. Kahraman, "Novel similarity measures in spherical fuzzy environment and their applications," *Engineering Applications of Artificial Intelligence*, vol. 94, pp. 1–15, 2020, doi: 10.1016/j.engappai.2020.103837.
- [55] M. Altaweel and A. Squitieri, "Quantifying object similarity: Applying locality sensitive hashing for comparing material culture," *Journal of Archaeological Science*, vol. 123, pp. 304–307, 2020, doi: 10.1016/j.jas.2020.105257.
- [56] N. Torkanfar and E. R. Azar, "Quantitative similarity assessment of construction projects using WBS-based metrics," *Advanced Engineering Informatics*, vol. 46, pp. 1–12, 2020, doi: 10.1016/j.aei.2020.101179.
- [57] D. Colla, E. Mensa, and D. P. Radicioni, "Novel metrics for computing semantic similarity with sense embeddings," *Knowledge-Based Systems*, vol. 206, pp. 1–15, 2020, doi: 10.1016/j.knosys.2020.106346.

BIOGRAPHIES OF AUTHORS






Riza Arifudin    received his education degree in Mathematics from Semarang State University (UNNES) Indonesia in 2003. He received his master's degree in computer science from Gadjah Mada University (UGM) Indonesia in 2010. Currently he is a candidate for Doctor of Information Systems, Graduate School, Diponegoro University, Indonesia. His research interests include learning management systems, mobile applications, and data mining. He can be contacted at email: rizaarifudin@mail.unnes.ac.id.



Subhan    received his education degree in Mathematics from Semarang State University (UNNES) Indonesia in 2012. He received his master's in Mathematics Education from Semarang State University (UNNES) Indonesia in 2015 and master's in Information Systems from Diponegoro University (UNDIP) Indonesia in 2016. Currently he is a lecturer in Department of Information Systems, FMIPA UNNES. His research interests include text mining, computer networks, and data mining. He can be contacted at email: subhan@mail.unnes.ac.id.



Yahya Nur Ifriza    received an education degree in Informatics and Computer Engineering from Semarang State University (UNNES) Indonesia in 2014. He received a master's degree in Information Systems from Diponegoro University (UNDIP) Indonesia in 2017. Currently he is a lecturer in the Department of Information Systems, FMIPA UNNES. His research interests include expert systems, internet of things, and data mining. He can be contacted at email: yahyanurifriza@mail.unnes.ac.id.