International Journal of Electrical and Computer Engineering (IJECE)
Vol. 10, No. 5, October 2020, pp. 4569~4580
ISSN: 2088-8708, DOI: 10.11591/ijece.v10i5.pp4569-4580

Data science for digital culture improvement in higher education using K-means clustering and text analytics

Dian Sa'adillah Maylawati[1], Tedi Priatna[2], Hamdan Sugilar[3], Muhammad Ali Ramdhani[4]
[1]Department of Informatics, UIN Sunan Gunung Djati Bandung, Indonesia
[1]Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka, Malaysia
[2]Department of Islamic Education, UIN Sunan Gunung Djati Bandung, Indonesia
[3]Department of Mathematics Education, UIN Sunan Gunung Djati Bandung, Indonesia
[4]Department of Informatics, UIN Sunan Gunung Djati Bandung, Indonesia

Article Info

Article history:

Received Apr 1, 2020
Revised Apr 13, 2020
Accepted Apr 23, 2020

Keywords:

Clustering
Data science
Digital culture
Higher education
K-means algorithm
Text analytics
Word cloud

ABSTRACT

This study aims to investigate the meaningful pattern that can be used to improve digital culture in higher education based on parameters of the technology acceptance model (TAM). The methodology used is the data mining technique with K-means algorithm and text analytics. The experiment using questionnaire data with 2887 respondents in Universitas Islam Negeri (UIN) Sunan Gunung Djati Bandung. The data analysis and clustering result show that the perceived usefulness and behavioral intention to use information systems are above the normal value, while the perceived ease of use and actual system use is quite low. Strengthened with text analytics, this research found that the EDA and K-means result in harmony with the hope or desire of academic society the information system implementation. This research also found how important the socialization and guidance of information systems, especially the new one information system, in order to improve digital culture in higher education.

Corresponding Author:

Dian Sa'adillah Maylawati,
Department of Infromatics, Faculty of Science and Technology,
UIN Sunan Gunung Djati Bandung,
A. H. Nasution Street, 105 Bandung, Indonesia
Centre for Advanced Computing Technology, Faculty of Information and Communication Technology,
Universiti Teknikal Malaysia Melaka,
Hang Tuah Jaya Street, 76100 Durian Tunggal, Melaka, Malaysia.
Email: diansm@uinsgd.ac.id

1. INTRODUCTION

In the technology disruption era, many software/applications/information systems built to help human activities. Tragically, 37% of 1,800 software are wasted, and 47% of them are software in the field of education [1]. This fact occurs because many factors such as unfulfilled user requirements; there are software errors, faults, and failures; software quality is not fulfilled; no innovation; does not apply the concept of human and computer interaction properly; difficult to use; not according to market needs (not up-to-date); to the lack of understanding of the use of technology due to its rapid development so that trends cannot be followed; etcetera. Of course, this is an obstacle for higher education to achieve *Techno University* [2], *Digital Campus* [3, 4], *Smart Campus* [5-8], *Green Campus* [9], as well as various other terms in the era of digital technology-based education. Therefore, no matter how sophisticated the technology is offered, when the application program is not used as planned, it will not have significant implications for human activity. One of the main problems of the failure of implementing digital systems is that they are not ready to accept technological changes so quickly that the use of technology is not cultivated and does not become a necessity

to support its activities. As a result, many applications are not used properly or rely on some people to use them. Digital culture in higher education must be established to support academic activities more effectively and efficiently. However, not all academicians in higher education aware about the benefits of digital literacy with various influence factors, such as lack of socialization, not easy to use, inadequate infrastructure, and so on. Seeing various issues related to the awareness, abilities, and culture of the academic community in using digital systems in higher education, a comprehensive and in-depth study of the factors influencing the use of digital systems is already in progress. The study can reveal the usability of software, both in terms of software/ applications and in terms of users.

Currently, data science is a popular technique to process data efficiently [10], with a big data character such as volume, variety, velocity, and so on [11-13]. Data science technique allows processing various types of data at once even in large amounts because data science combines the statistical process with data mining or machine learning. Where data mining is a computational technique to find insight knowledge from large data [14]. In data science, there is the *exploratory data analysis* (EDA) technique that prepares and processes the data statistically [15]. Statistical data processing techniques, known as exploratory data analysis (EDA), are considered capable of preparing data before it is processed properly. EDA can reduce redundant data, is considered not to affect the results, complete the missing value, complete the data, so that other things that maximize the data to be cleaner are carried out later. Currently, EDA is also developing using prediction models and machine learning [16]. One of the most popular machine learning/ data mining algorithms that can be used in the EDA process is the K-means clustering algorithm [17-21].

There are several related previous kinds of research that use data science and data mining technique to process data, among others: (1) data science and harnessing analytics used to get a meaningful assessment for learning activities [22]; (2) educational data science was used to evaluate students' usage in the massive open online course [23]; (3) data science approach also used for identifying the crucial factors for assessment of an international student with predicting the exam result [24]; (4) data mining approach with K-means and Naïve Bayes algorithm also used for understanding the digital learning sources [25]; (5) K-means algorithm was proven as accurate evaluation for learning evaluation based on brainwave-based emotion [26]; (6) there is research that evaluates teacher's experience in digital content evaluation using qualitative thematic analysis with K-means cluster analysis [27]; and (7) the learning behavior pattern of the digital textbook was analyzed using clustering method using K-means algorithm [28]. Based on many previous kinds of research that use data science with data mining, especially the K-means clustering algorithm, to evaluate digital learning, this study aims to investigate, interpret, and find the meaningful pattern of digital culture in higher education using K-means algorithm. The interpretation result can be a meaningful knowledge in responding, developing, and improving digital culture in higher education.

2. RESEARCH METHOD

2.1. Research activities

The case of this research is UIN Sunan Gunung Djati Bandung that is one of the higher education with a vision to become a superior and competitive campus through the use of technology. No less than 58 information systems in the UIN Sunan Gunung Djati Bandung environment that support all education activities, ranging from admissions, Academic services administration systems, Financial information systems, employee information systems, e-Library, e-Learning, systems registration of assemblies, Helpdesk Systems, and various information systems and other applications. However, it turns out that awareness and the need for the use of the information system provided are not evenly distributed throughout the academic community, there are still those who rely on each other, even indifferent.

The research activity depicted in Figure 1 starts from literature studies related to data science, data mining, and the K-means algorithm, which then compiles questions for questionnaires to be distributed to stakeholders in various faculties, study programs/ departments, to units in the environment. UIN Sunan Gunung Djati Bandung. The questionnaire data is then processed using EDA and clustering the K-means algorithm. The results of EDA and K-means data processing are analyzed, studied, and interpreted to find a meaningful pattern so that can produce a recommendation model for strategies to strengthen digital culture in the academic community of Sunan Gunung Djati University, Bandung. And this research use Python as programming language for EDA, K-means, and text analytics [29].

The experiment of this study utilized the Google colaboratory (Google colab) with Python as programming language for EDA and text analytics. For the K-means clustering, this study used Orange as data mining tools. The visualization of EDA, K-means clustering, and text analytics is provided by the Python and Orange library.

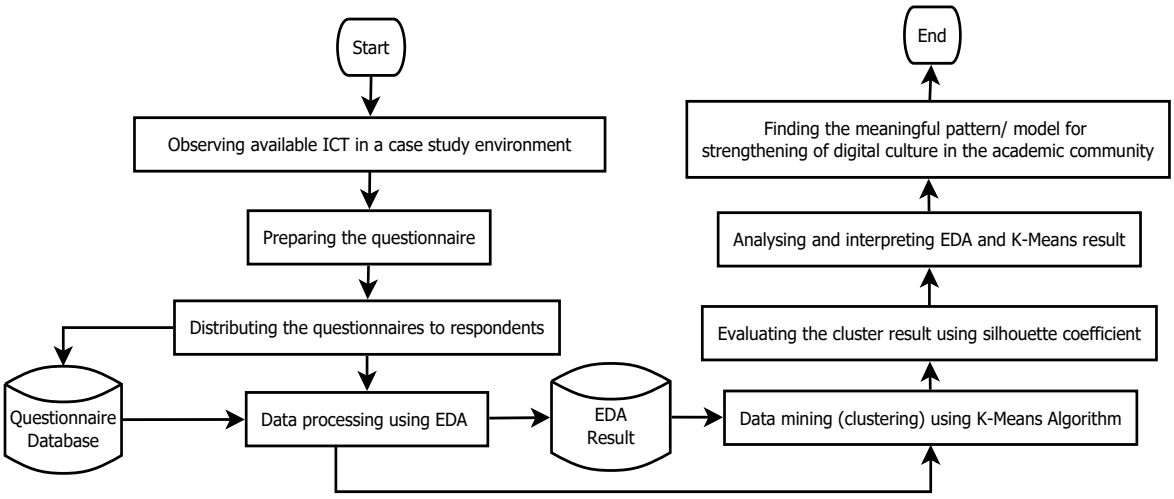

Figure 1. Research activities

2.2. Data collecting

Data collection was carried out by distributing questionnaires using Google forms with a total of 60 questions distributed to the ranks of the Rectorate, Dean, Senate, Bureau, Institution, and SPI (as policymakers), 9 Faculties and Postgraduate (involving students, lecturers, and education staff from 15 study programs in Postgraduate, 5 majors in the Faculty of *Ushuluddin*, 5 majors in the Faculty of *Tarbiyah* and Teacher Training, 6 study programs in the Faculty of Sharia and Law, 4 majors in the Faculty of *Da'wah* and Communication, 3 majors in the Faculty of *Adab* and Humanities, 1 department in the Faculty Psychology, 7 majors in the Faculty of Science and Technology, 3 majors in the Faculty of Social and Political Sciences, and 4 study programs in the Faculty of Economics and Islamic Business), 11 Technical Service Units, 5 General and Mahad Service Units. The questionnaire was prepared with the concept of the Technology Acceptance Model, among others: perceived ease of use, perceived usefulness, behavior intention to use, and actual system use.

2.3. Exploratory data analysis

Exploratory Data Analysis is a procedure to analyze data easily, accurate, precise with mathematical statistics as an output, where the process is automatically by machine [30, 31]. Basically, EDA provides a summary of numerical data such as average, median, maximum value, minimum value, and quartile. EDA aims to suggest hypotheses about the causes of observed phenomena, to assess assumptions on which to base statistical conclusions, to support the selection of appropriate statistical techniques, and to provide a basis for further data collection. EDA results are usually visualized using graphical techniques, such as square plots, histograms, Pareto diagrams, distribution plots, multidimensional scaling, principal component analysis, and interactive version of the plot. In data mining or machine learning techniques, EDA is usually used in the pre-processing process to visualize, find missing, and also to look for correlations between data or variables. Because the pre-processing phase is important for data selection, data cleaning to improve quality, data transformation, and data reduction to run an efficient mining process.

2.4. K-means algorithm

Data mining is a technique for finding important information or insight knowledge from big data [32]. Where, data mining has four main approaches, among others, classification (classification) which is supervised learning, clustering which is unsupervised learning, association rule, and semi-supervised learning that combines classification and clustering. Data mining is used to find hidden information that is important and can be used to predict and support decision making. Clustering is not used to predict like classification, but clustering will produce the insight of data that problematic and analyzed and interpreted by human [19, 33]. K-means is one of the most widely used clustering algorithms that find minimum distance values in the same cluster [34-36]. K-means is a simple algorithm with fast processing time and produces an optimal cluster. The K-means algorithm is as follows:

1. Determine the number of clusters.
2. Initiate the centroid value for each cluster ($\mu_1, \mu_2, \ldots, \mu_K \in \mathbb{R}$) randomly.
3. Repeat the calculation with the formula (1) and (2) until convergent.

4. For each i, calculate:

$$c^{(i)} := \arg min_j \left\| x^{(i)} - \mu_j \right\|^2 \ldots \quad (1)$$

5. And for each j, calculate:

$$\mu_j := \frac{\sum_{i=1}^{m} 1\{c^{(i)}=j\} x^{(i)}}{\sum_{i=1}^{m} 1\{c^{(i)}=j\}} \ldots \quad (2)$$

2.5. Text analytics

Text analytics is a technique to find meaningful knowledge from text data [37]. Text analytics is not always called text mining, because text mining always contains the mining process inside it, such as classification, clustering, or association rule for text data. But, text mining is a part of text analytics also. Another text analytics technique such as sentiment analysis [38-40], opinion mining [41], social media analysis [42-45], social networks [46], and web scraping and crawling [47]. Several literature said that text analytics is a part of natural language processing (NLP) [48], because not all NLP use text data as language database, it can be voice/ sound, image, and video. Information retrieval [49], semantic search engine [50], text similarity or string matching [51], and text summarization [52] are a type of NLP that commonly uses text data.

2.6. Silhouette coefficient

The clustering result should be measured to ensure that the resulting pattern is good enough. There are internal and external measurements. External measurements like Jaccard Coefficient [53], Purity [54], Precision and Recall [55], F-Measure [56], and so on. Whereas, internal measurements such as Z-Score Index [57], Gamma and Somer's Gamma [58], Silhouette coefficient [59], BetaCV and Dunn index [60], and so on. Silhouette coefficient is widely used to evaluate the results of clustering. Silhouette coefficient is a metric that measures cluster separation and compactness at the same time [59, 61-63]. Formula (3) is used to calculate the average distance in a cluster and the minimum distance between objects to other clusters,

$$SC = \frac{1}{N} \sum_{i=1}^{N} \frac{\beta_i - \alpha_i}{max\{\alpha_i, \beta_a\}'} \quad (3)$$

where, α_i is the average distance of objects in a cluster, i.e. (formula (4)):

$$\alpha_i = \frac{\sum j \neq 1, x_j \in c_i |X_i - X_j|}{|C_i|} \quad (4)$$

and β_i is a distance between the object x_i with nearest centroid center w_j. β_i calculated by the formula (5):

$$\beta_i = \min\{|x_i - w_j|, j = 1,2, \ldots, k, j \neq 1\} \quad (5)$$

Silhouette Coefficient values range between 1 to -1 ($-1 \leq Silhouette\ Coefficient \leq 1$), where 1 means the grouping solution is "correct" and -1 means the grouping solution is "wrong". However, according to the results of clustering, it does not offer a guarantee of accuracy, but many interpretations of the results of clustering. So, there is no guarantee that the Silhouette Coefficient value close to 1 always has the right cluster and many interpretations, and vice versa.

3. RESULTS AND DISCUSSIONS

3.1. Data collection

Data collection which was successfully obtained is a totally of 2887 data from 338 Lecturers, 200 Educational Personnel, and 2349 Students of UIN Sunan Gunung Djati Bandung. This data already fulfilled 10% of the population. However, to meet the quality of the result, the missing value and outlier are decided to be deleted. So, total data that used are 2365 respondent data with 298 Lecturer, 128 Educational Personnel, and 1939 Student. While the total female is 1348 respondents and 1071 respondents are male. The questions are collected based on parameters of the technology acceptance model (TAM), such as perceived usefulness (PU), Perceived ease of use (PEU), Behavioural intention to use (BIU), and Actual system use (ASU) [64, 65]. TAM also already used to evaluate information technology in higher education, such as e-learning or learning management systems [66, 67]. This research has 6 questions related to PU, 11 questions for PEU, 10 questions for BIU, and 12 questions for ASU.

3.2. Result of exploratory data analysis and data clustering

3.2.1. Exploratory data analysis result

The result of exploratory data analysis (EDA) shows several conclusions related to the implementation of information technology in UIN Sunan Gunung Djati Bandung and can be used as a basis for enhancing digital culture in higher education. This analysis is based on the result of EDA and the correlation between parameters that visualized in Figure 2. The analysis results, among others:

a. Overall, the perceived usefulness of information systems is above average with value is 3.43. Perceived usefulness indicates the level of confidence in individuals that technology can improve their performance [68]. Therefore, the result value of perceived usefulness in UIN Sunan Gunung Djati Bandung can be concluded that academic society aware of the usefulness of digital technology that can make their activities more efficient and effective. This value is supported by 65.62% respondents know about the term of the information system, 69% knows about the benefits of information technology/ information system, 62.41% respondents know the main website that provides the up-to-date information about academic activities and news. However, the information sharing about the information system is low, only 42%. This fact shows that today, the academic society has been aware that their academic activities can not be separated from technology support. Therefore, to improve digital culture in higher education, the technology not always used every time but the perception of the usefulness of technology must always be maintained [69]. This can be realized if the technology that available support the needs of users, so that the usefulness can be felt.

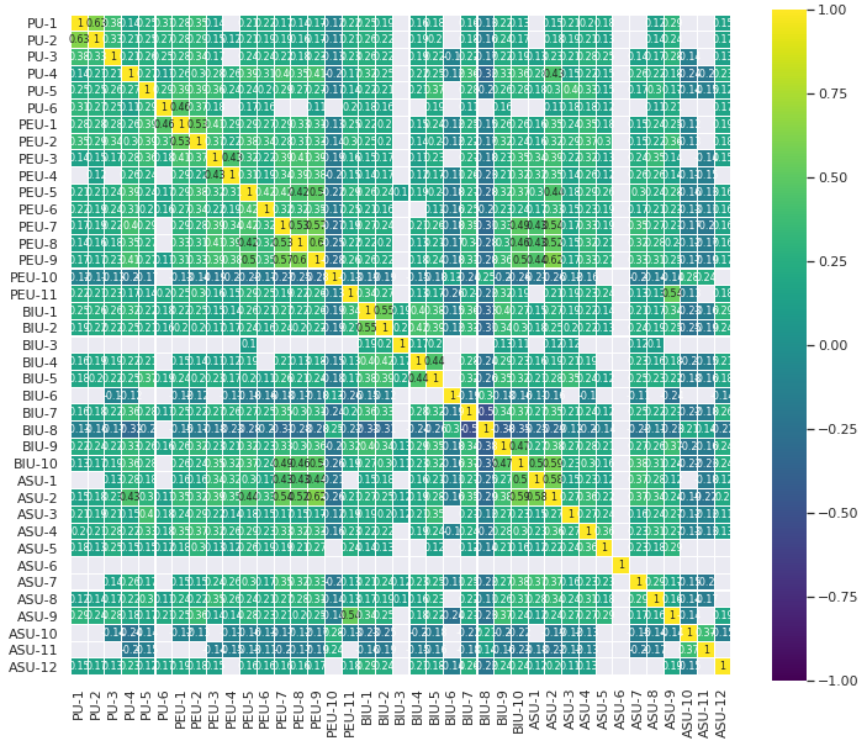

Figure 2. Correlation between TAM parameters

b. The perceived ease of use of the information system which available in UIN Sunan Gunung Djati Bandung is quite low, below the normal value, it is 2.67. Perceived ease of use indicates the use of an information system whether easy to use or understand [70]. The result value means that many respondents feel difficulty in using the system. The fact shows that (31.07% respondent) is only 25% of all information system which provide the manual guide because of only a half of system that provides the complete manual guide and conducted the socialization or training how to use the system. Then, 42.16% of respondent agree that half of the information system easy to use, 44.95% respondent agree that half of system user interface is interesting so easy to understand, and 45.92% respondent feel that only a half of system that fulfills the requirements of the process business through the functions that available. Most of them decide to use the system although they need more time for understanding

the system, even if there is a new system that released (58.56%). This result concludes that the socialization, training, and manual guide for the system is important to improve the perceived ease of use of the information system in higher education. The impact is the digital culture in higher education will be improved too.

c. The behavior of intention is defined as an effort or strong desire from the users to try and use the system [71]. In UIN Sunan Gunung Djati Bandung, the behavior of intention to use the information system is above the normal value, around 3.28. It means that digital culture in UIN Sunan Gunung Djati Bandung already awakened. It is proven by almost 50% of respondents' support and enthusiasm to try and use the new information system if it is released. Because 52.26% of respondent feel that using the information system can support their academic activities efficiently and effectively. This understanding must be continually developed and maintained to improve digital culture in higher education, because the behavioral intention to use information technology can be the habit of the digital user [72], especially academic society in using academic information systems such as the learning management system [73].

d. For the whole information system that available in UIN Sunan Gunung Djati Bandung, the actual system uses is still quite low below the normal value, it is 2.65. This value means in the implementation of the information systems is still low, such as 44.27% of respondents agree that only half of the system that fulfills the user or business process requirements, only 29.69% of respondents that remember the link address to access the system that they need. Even though, 78.14% of respondents decide to use a digital system to support their academic activities. This result proves that it is important to design the system that involving users in its development to fulfill their requirements. Because every type of user has a different requirement that must be accommodated, analyzed, and selected so as to best meet the needs of all users. The good information system/ software design will produce a good quality of information system [74], then the digital culture will be improved if all they need are accommodated.

3.2.2. Result and interpretation of the K-means algorithm

Figures 3-5 visualized the result of the K-means algorithm in clustering the type of data (the clustering result and the example of a cluster member). The clusters are formed due to the similarity of data characteristics. Based on the silhouette coefficient value with Euclidean distance, the best cluster for this research data is two clusters, with 1124 in Cluster 1 (C1 - Blue) and 1241 in Cluster 2 (C2 - Red). Therefore, Figure 3 visualize the result with two cluster-based on Respondent (Lecturer = 1; Educational Personnel = 2; Student = 3), Figure 4 is based on Gender (Male = 1; Female = 2), while Figure 5 based on Age. The initiation of the centroid center is assigned randomly. K-means algorithm uses random initialization with 300 maximum iterations.

Actually, the K-means clustering result is not reliable enough, there are many members of the cluster (both blue cluster and red cluster) that far apart from the centroid. Many members that also have the similarity, whereas they are in a different cluster. The Silhouette coefficient value of this cluster is too low (0.125) so that the cluster is quite difficult to be interpreted. However, the cluster region/ area is still quite clearly separated (visualized with a background color).

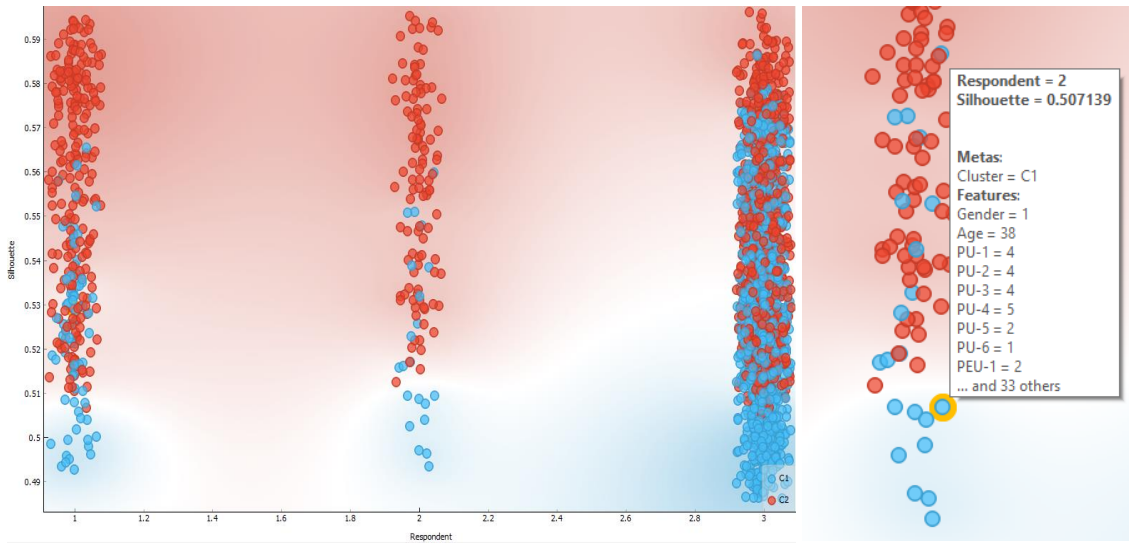

Figure 3. Cluster result based on respondent

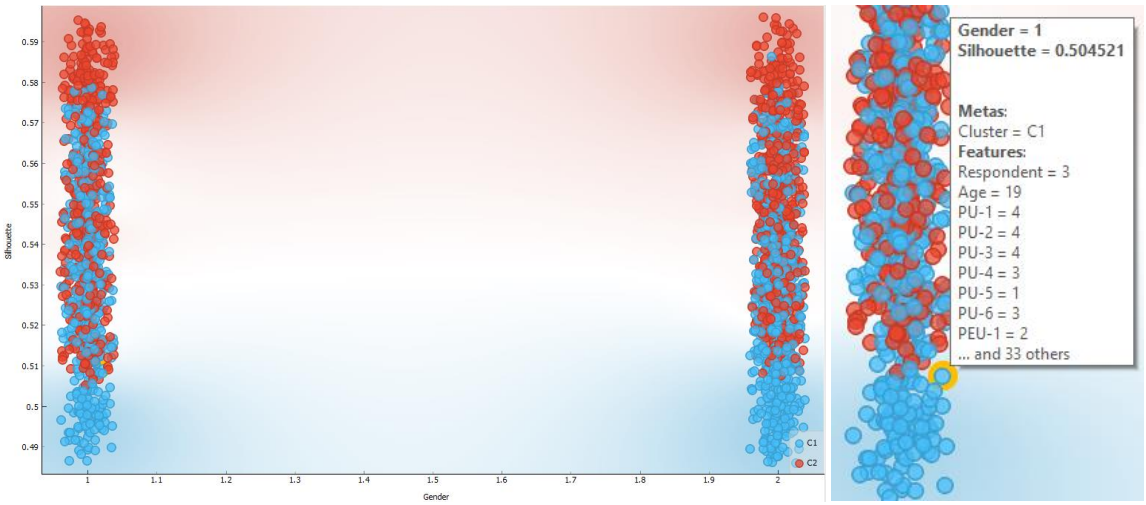

Figure 4. Cluster result based on gender

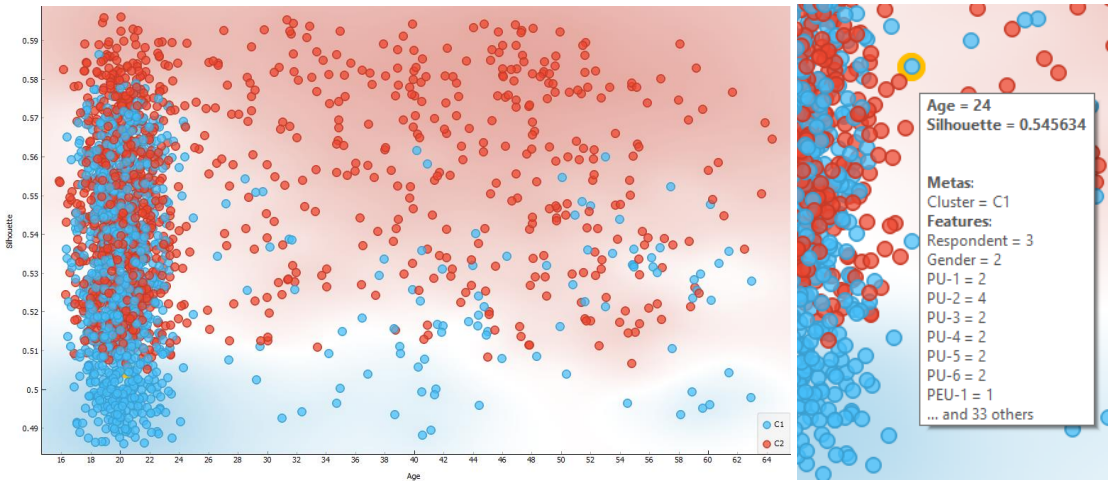

Figure 5. Cluster result based on age

When examined further, there are several conclusions that can be obtained, among others:

a. The C1 is a group of respondents that have perceived ease of use at the lowest level (under 2), it means that 47.53% of respondents still feel need the more effort to use/ adapt/ learn the information system because of lack of the information and socialization of system. On the other hand, 56.47% of the respondent (C2) feel normal or even feel easy in use the system. But, C2 has the highest perceived usefulness. It means that most of the respondents know and understand the benefits of digital technology to support academic activities in higher education.
b. Based on the clustering result in Figure 3, compared with the student, most of the lecturer and educational personnel are in C2. It means that the lecturer and educational personnel can learn the system easier that student. This fact shows that the socialization has not been evenly distributed, it should be assumed that socialization/ training on the use of the system is mostly carried out by lecturers and educational personnel than the student.
c. Based on gender in Figure 4, there is no significant cluster difference. Moreover, there are too many members of C1 and C2 that in the same cluster region. However, it is shown that gender does not affect the use of technology and digital culture in higher education.
d. Because of the high age variation (visualized in Figure 5), it appears that the clusters formed are not compact. However, what can be obtained from cluster results based on age that in the age range above 30 years, more in C2. This shows that they can use the information system well and feel the benefit of the information system because of the support of good socialization and training systems.

3.3. Result of text analytics

3.3.1. Test pre-processing

Text pre-processing is an important phase in text analytics, including in text mining and natural language processing. Text pre-processing is a phase to prepare text data until ready to process in the next phase and ensure the quality of text data [75, 76], either in the input process or result process. Not all process in the text pre-processing is used, sometimes it is in accordance with the needs of the research. Generally, there are tokenizing, lower case (case folding), remove regular expression, stop-word removing, and stemming process in the text pre-processing process [77]. Every language has different characteristics, structures, and grammar, including the Indonesian language. This research uses the Indonesian language. The text data is collected is contained the message, impression, and hope from 2887 respondents.

3.3.2. Result and interpretation of text analytics

Figure 6 illustrated the word cloud based on the frequency of words. As shown in Figure 7 top 15 of words that appeared from text data are: *lebih* (more), digital, *aplikasi* (application), *mahasiswa* (college student), *system* (system), *semoga* (hope/ wish), *baik* (good/ well), UIN, online, *yg* (abbreviation of *yang*-preposition in Indonesian language), *banyak* (a lot of/ many/ much), *tidak* (no), *nya* (possessive pronoun in Indonesian language), *sosialisai* (socialization), and *information* (informasi). Actually, the word such as *yg, tidak, nya, tidak,* and many more which are abbreviated and included in the stop-word category, unsuccessfully removed. And also, several affixes in the stemming process is not changed. It happens because this experiment uses the Sastrawi library for Python without improvement for this case [78].

Based on the text analytics result about the message and hope of respondent including lecturer, educational personnel, and student, it can be concluded that:

a. In accordance with the result of perceived ease of use (PEU) and actual system use (ASE) which are low, the respondent (especially student) hopes that socialization or training of information system must be comprehensive and massive. This can improve the digital culture in higher education that introduces the system (or new system) completely and thoroughly for all end-users. Not only certain groups, because each user has a different level of understanding and adjustment about the information system. The digital natives who born more than 1980 and familiar with digital technology allegedly faster in understanding new systems or technology than immigrants natives [79].
b. The socialization must be supported by a manual book which completes for each information system. The fact of the survey shows that manual books not all available for each system, and also incomplete instructions for use in the system. Even though the manual book is available, but it has not shared/ inform/ socialized well, so that not all end-user get the manual book or can search the manual book easily. Especially for the student, in accordance with the K-means clustering result, most of the students are in C1 who feel need more effort to use the system because of the lack of socialization.
c. Generally, respondents hope that the information system, *budaya digital/* digital culture, *system digital/* digital system, online system in UIN Sunan Gunung Djati Bandung *lebih baik*/ better than before. This hope proves that digital culture in UIN Sunan Gunung Djati Bandung already awakened. It is in accordance with BIU results that above average, interest, support, and desire to use information technology are quite high. This needs to be supported by system functions that meet the needs of academic society, thorough socialization, and a complete manual book. So that digital culture will further improve because the system is easy to use, according to the needs of academic society, and has a direct impact on performance because work becomes more effective and efficient.

Figure 6. Word cloud result

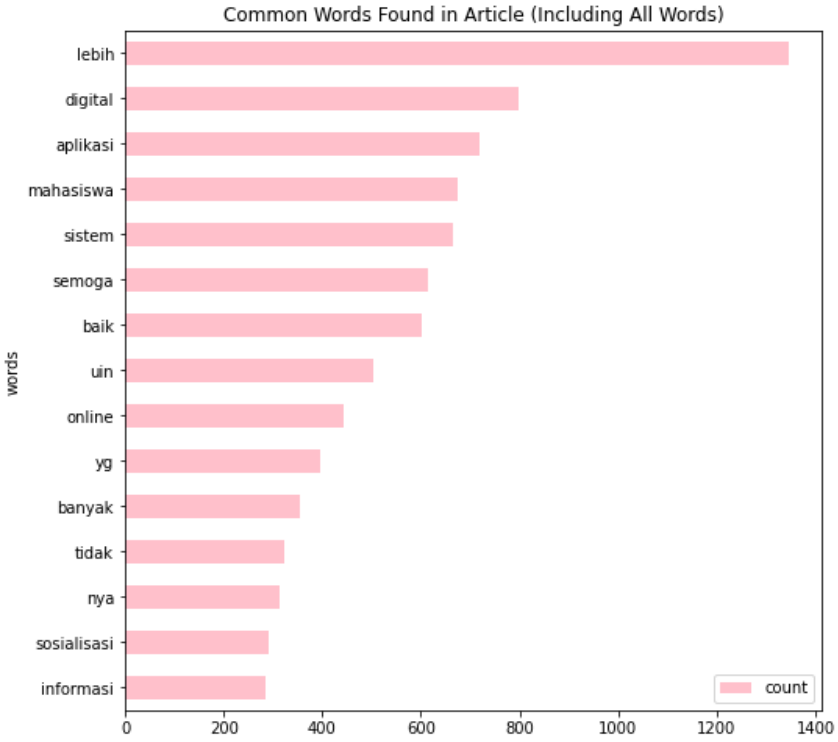

Figure 7. Top 15 frequent of words

4. CONCLUSION

This research is conducted comprehensively in order to evaluate the information technology implementation in higher education. The questionnaire data from lecturers, students, and educational personnel are analyzed using exploratory data analysis (EDA), K-mean clustering algorithm, and also text analytics. The result of the experiment of EDA and K-means algorithm shows that to improve the digital culture with high perceived ease of use and actual system use of information technology should be supported with complete and comprehensive socialization, and also provide the manual guide for each information system. This result in accordance with the hope of end-user that need information, knowledge, and guideline for an information system that they used. Digital culture through behavioral intention use of information system that already awakened should be maintained and improve with the quality of information system which fulfills the user requirements.

For further works, it needs to prepare the data better so that it can produce the reliable cluster, although clustering is not used to predict, it can produce a more accurate interpretation if the data prepared is better. The other clustering methods can be used to get a better cluster. And also, it can use the classification approach to predict the type of respondent and the result can be used as decision support by policymaker in higher education related to information technology and digital culture improvement.

ACKNOWLEDGEMENTS

The researchers would like to appreciate and many thanks to Rector and the academic society of UIN Sunan Gunung Djati Bandung who support this research.

REFERENCES

[1] 1E Report, "The Real Cost of Unused Software," *1E Company*, 2015.
[2] D. Jamaluddin, M. A. Ramdhani, T. Priatna, and W. Darmalaksana, "Techno University to increase the quality of islamic higher education in Indonesia," *International Journal of Civil Engineering and Technology*, vol. 10, no. 1, pp. 1264-1273, 2019.
[3] D. Kurniadi, "Perancangan Arsitektur Sistem E-academic dengan Konsep Kampus Digital Menggunakan Unified Software Development Process (USDP)-Architectural Design of E-academic Systems with Digital Campus Concepts Using the Unified Software Development Process (USDP)," *Jurnal Wawasan Ilmiah Manajemen dan Teknik Informatika*, vol. 5, no. 10, Mar. 2014.

[4] M. E. Mahmud, "Mewujudkan Sekolah Atau Kampus Digital-Creating a School or Campus Digital," *Dinamika Ilmu*, vol. 1, no 1, 2011. DOI: doi.org/10.21093/di.v11i1.46

[5] A. Abuarqoub *et al.*, "A survey on internet of things enabled smart campus applications," in *ACM International Conference on Future Networks and Distributed Systems (ICFNDS 2017)*, Cambridge, United Kingdom, pp. 1-7, 2017. DOI: doi.org/10.1145/3102304.3109810

[6] X. Dong, X. Kong, F. Zhang, Z. Chen, and J. Kang, "OnCampus: a mobile platform towards a smart campus," *Springerplus*, vol. 5, no. 974, 2016. DOI: doi.org/10.1186/s40064-016-2608-4

[7] M. R. Veeramanickam and M. Mohanapriya, "IOT enabled Futures Smart Campus with effective E-Learning : i-Campus," *Journal of Engineering Technology (JET)*, vol. 3, no. 4, pp. 81-87, April 2016.

[8] A. Alghamdi and S. Shetty, "Survey toward a smart campus using the internet of things," in *Proceedings - 2016 IEEE 4th International Conference on Future Internet of Things and Cloud (FiCloud)*, Vienna, pp. 235-239, 2016.

[9] H. I. Wang, "Constructing the green campus within the internet of things architecture," *Int. J. Distrib. Sens. Networks*, vol. 10, no. 3, pp. 1-8, 2014.

[10] B. Manoj, K. V. K. Sasikanth, M. V. Subbarao, and V. Jyothi Prakash, "Analysis of data science with the use of big data," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 7, no. 6, pp. 87-90, 2018.

[11] P. S. Arockia, S. S. Varnekha, and K. A. Veneshia, "The 17 V's of Big Data," *Int. Res. J. Eng. Technol.*, vol. 4, no. 9, pp. 3-6, 2017.

[12] S. Sagiroglu and D. Sinanc, "Big data: A review," in *Proceedings of the 2013 International Conference on Collaboration Technologies and Systems, CTS 2013*, San Diego, CA, pp. 42-47, 2013.

[13] K. Borne, "Top 10 List – The V's of Big Data," *Data Science Central*, 2014. [Online]. Available: https://www.datasciencecentral.com/profiles/blogs/top-10-list-the-v-s-of-big-data.

[14] H. Jiawei, M. Kamber, J. Han, M. Kamber, and J. Pei, "Data Mining: Concepts and Techniques," *Morgan Kaufmann*, 2nd Ed., 2006.

[15] M. Abzalov, "Exploratory data analysis," in *Modern Approaches in Solid Earth Sciences*, 2016.

[16] G. Saporta, "50 Years of Data Analysis: From Exploratory Data Analysis to Predictive Modeling and Machine Learning," Data Anal. Appl. 1 Clust. Regression, Model. Forecast. Data Min., ISTE-Wiley, Data Analysis and Applications, 978-1-78630-382-0. ffhal-02470740f, 2019.

[17] A. Kazemi and G. Khodabandehlouie, "A new initialisation method for k-means algorithm in the clustering problem: data analysis," *Int. J. Data Anal. Tech. Strateg.*, vol. 10, no. 3, pp. 291-304, 2018.

[18] I. D. Dinov, "K-Means Clustering," in *Data Science and Predictive Analytics*, Springer, pp. 443-473, 2018.

[19] P. Nerurkar, A. Shirke, M. Chandane, and S. Bhirud, "Empirical analysis of data clustering algorithms," *Procedia Comput. Sci.*, vol. 125, pp. 770-779, 2018.

[20] H. Anderson and G. Ascoli, "Exploratory Data Analysis of Autobiographical Memory Trends," *J. Student-Scientists' Res.*, vol. 1, 2019.

[21] J. T. Chi, E. C. Chi, and R. G. Baraniuk, "K-pod: A method for K-means clustering of missing data," *Am. Stat.*, vol. 70, no. 1, pp. 91-99, 2016.

[22] D. Ifenthaler, S. Greiff, and D. Gibson, "Making use of data for assessments: Harnessing analytics and data science," in *International handbook of IT in primary and secondary education (2 ed.)*, Springer, 2018.

[23] C. Romero and S. Ventura, "Educational data science in massive open online courses," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 7, no. 1, pp. e1187, 2017.

[24] L. Gaid and C. H. Yu, "A data science approach to identify crucial factors of predicting test performance in Program for International Student Assessment," *SCCUR Southern California Conferences for Undergraduate Research*, 2019.

[25] H. Praherdhiono, E. P. Adi, and R. N. Devita, "Understanding of Digital Learning Sources with the Heutagogy Approach using the K-Means and Naive Bayes Methods," in *2018 4th International Conference on Education and Technology (ICET)*, pp. 23-27, 2018.

[26] T. M. Li, H. H. Cho, H. C. Chao, T. K. Shih, and C. F. Lai, "An accurate brainwave-based emotion clustering for learning evaluation," in *International Symposium on Emerging Technologies for Education*, pp. 223-233, 2017.

[27] M. K. Kim, K. Xie, and S.-L. Cheng, "Building teacher competency for digital content evaluation," *Teach. Teach. Educ.*, vol. 66, pp. 309-324, 2017.

[28] C. Yin, Z. Ren, A. Polyzou, and Y. Wang, "Learning Behavioral Pattern Analysis Based on Digital Textbook Reading Logs," in *International Conference on Human-Computer Interaction*, pp. 471-480, 2019.

[29] D. Sarkar, "Text Analytics with Python - A Practical Real-World Approach to Gaining Actionable Insights from Your Data," *APress*, 2016.

[30] J. W. Tukey, "The Future of Data Analysis," *Ann. Math. Stat.*, vol. 33, no. 1, pp. 1-67, 1962.

[31] D. C. Hoaglin, "John W. Tukey and Data Analysis," *Statistical Science*, vol. 18, no. 3, pp. 311-318, 2004.

[32] H. Jiawei, M. Kamber, J. Han, M. Kamber, and J. Pei, "Data Mining: Concepts and Techniques," *Morgan Kaufmann, 3rd Ed.*, 2012.

[33] P. Nerurkar, A. Shirke, M. Chandane, and S. Bhirud, "A novel heuristic for evolutionary clustering," *Procedia Comput. Sci.*, vol. 125, pp. 780-789, 2018.

[34] J. A. Hartigan and M. A. Wong, "Algorithm AS 136: A K-Means Clustering Algorithm," *Journal of the Royal Statistical Society. Series C, (Applied Statistics)*, vol. 28, no. 1, pp. 100-108, 1979.

[35] C. Slamet, A. Rahman, M. A. Ramdhani, and W. Dharmalaksana, "Clustering the verses of the holy qur'an using K-means algorithm," *Asian Journal of Information Technology*, vol. 15, no. 24, pp. 5159-5162, 2016

[36] A. D. Rachid, A. Abdellah, B. Belaid, and L. Rachid, "Clustering Prediction Techniques in Defining and Predicting Customers Defection: The Case of E-Commerce Context," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 8, no. 4, pp. 2367-2383, 2018.
[37] V. Kagan, E. Rossini, and D. Sapounas, "Sentiment Analysis for PTSD Signals," in *SpringerBriefs in Computer Science*, 2013.
[38] Aylien, "Aylien Text Analytics," [Online]. Available: https://aylien.com/.
[39] Rosette, "Rosette Text Analytics," [Online]. Available: https://www.rosette.com/.
[40] D. M. E. D. M. Hussein, "A survey on sentiment analysis challenges," *J. King Saud Univ. - Eng. Sci.*, vol. 30, no. 4, pp. 330-338, 2018.
[41] R. Bhalla and A. Bagga, "Opinion mining framework using proposed rb-bayes model for text classification," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 9, no. 1, pp. 477-485, 2019.
[42] S. Chen, L. Lin, and X. Yuan, "Social Media Visual Analytics," *Comput. Graph. Forum*, vol. 36, no. 3, pp. 563-587 2017.
[43] I. Lee, "Social media analytics for enterprises: Typology, methods and processes," *Bus. Horiz.*, vol. 61, no. 2, 2018.
[44] L. Branz and P. Brockmann, "Sentiment Analysis of Twitter Data," in *Proceedings of the 12th ACM International Conference on Distributed and Event-based Systems - DEBS '18*, 2018.
[45] A. Ali, et al., "Sentiment Analysis on Twitter Data using KNN and SVM," *International Journal of Advanced Computer Science and Applications(IJACSA)*, vol. 8, no. 6, pp. 19–25, 2017.
[46] P. Nerurkar, et al., "A comparative analysis of community detection algorithms on social networks," in *Computational Intelligence: Theories, Applications and Future Directions*, Springer, vol. 1, pp. 287–298, 2019.
[47] C. Slamet, et al., "Web Scraping and Naïve Bayes Classification for Job Search Engine," *IOP Conference Series: Materials Science and Engineering*, vol. 288, no. 1, pp. 1–7, 2018.
[48] P. M. Nadkarni, L. Ohno-Machado, and W. W. Chapman, "Natural language processing: An introduction," *J. Am. Med. Inform. Assoc.*, vol. 18, no. 5, pp. 544–551 2011.
[49] V. V. Raghavan, V. N. Gudivada, Z. Wu, and W. I. Grosky, "Information retrieval," in *The Practical Handbook of Internet Computing*, 2004.
[50] J. M. Kassim and M. Rahmany, "Introduction to semantic search engine," in *Proceedings of the 2009 International Conference on Electrical Engineering and Informatics, ICEEI 2009*, pp. 380-386, 2009.
[51] W. H. Gomaa and A. A. Fahmy, "A Survey of Text Similarity Approaches," *Int. J. Comput. Appl.*, vol. 68, no. 13, pp.13-18, 2013.
[52] M. Allahyari *et al.*, "Text Summarization Techniques: A Brief Survey," *Int. J. Adv. Comput. Sci. Appl.*, vol. 8, no. 10, pp. 397-405 2017.
[53] J. Santisteban and J. L. Tejada Carcamo, "Unilateral Jaccard similarity coefficient," in *CEUR Workshop Proceedings*, 2015.
[54] S. C. Sripada and M. Sreenivasa Rao "Comparison of Purity and Entropy of K-Means Clustering and Fuzzy C Means Clustering," *Indian J. Comput. Sci. Eng.*, vol. 2 no. 3, pp. 343-346, 2011.
[55] Jumadi, D. S. A. Maylawati, B. Subaeki, and T. Ridwan, "Opinion mining on Twitter microblogging using Support Vector Machine: Public opinion about State Islamic University of Bandung," in *Proceedings of 2016 4th International Conference on Cyber and IT Service Management, CITSM 2016*, 2016.
[56] M. Azhari and Y. Jaya Kumar, "Improving text summarization using neuro-fuzzy approach," *J. Inf. Telecommun.*, vol. 1, no. 4, pp. 367–379, 2017.
[57] V. Kathiresan and P. Sumathi, "An efficient clustering algorithm based on Z-Score ranking method," *Proc. 2012 Int. Conf. Comput. Commun. Informatics*, pp. 1–4, 2012.
[58] E. H. Ahmad *et al.*, "Relationship of Work Stress to the Performance of Intensive Care Unit Nurses in Makassar," *Am. J. Public Heal. Res.*, vol. 6, no. 1, pp. 18-20 2018.
[59] P. Bafna, D. Pramod, and A. Vaidya, "Document clustering: TF-IDF approach," in *International Conference on Electrical, Electronics, and Optimization Techniques, ICEEOT 2016*, pp. 61-66, 2016.
[60] P. Nerurkar, A. Pavate, M. Shah, and S. Jacob, "Performance of internal cluster validations measures for evolutionary clustering," in *Computing, Communication and Signal Processing*, Springer, pp. 305-312, 2019.
[61] M. Anggara, H. Sujiani, and N. Helfi, "Selection of Distance Measure in K-Means Clustering for Grouping Members at Alvaro Fitness (in Bahasa)," *J. Sist. dan Teknol. Inf.*, vol. 1, no. 1, pp. 1-6, 2016.
[62] S. Aranganayagi and K. Thangavel, "Clustering categorical data using silhouette coefficient as a relocating measure," in *Proceedings - International Conference on Computational Intelligence and Multimedia Applications, ICCIMA 2007*, pp. 13-17, 2008.
[63] S. Luan, X. Kong, B. Wang, Y. Guo, X. You, "Silhouette coefficient based approach on cell-phone classification for unknown source images," in *IEEE International Conference on Communications*, pp. 6744-6747, 2012.
[64] M. McCord, "Technology acceptance model," in *Handbook of Research on Electronic Surveys and Measurements*, 2006.
[65] P. Surendran, "Technology Acceptance Model : A Survey of Literature," *Int. J. Bus. Soc. Res.*, vol. 2, no. 4, pp. 175-178, 2012.
[66] A. Q. M. AlHamad, "Acceptance of E-learning among university students in UAE: A practical study," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 10, no. 4, pp. 3660-3671, 2020.
[67] M. K. Alsmadi, "The students ' acceptance of learning management systems in Saudi Arabian Universities," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 10, no. 4, pp. 4155–4161, 2020.
[68] F. D. Davis, "User acceptance of information technology: system characteristics, user perceptions and behavioral impacts," *Int. J. Man. Mach. Stud.*, vol. 38, no. 3, pp. 475-487, 1993.

[69] M. Henderson, N. Selwyn, G. Finger, and R. Aston, "Students' everyday engagement with digital technology in university: exploring patterns of use and 'usefulness," *J. High. Educ. Policy Manag.*, vol. 37, no. 3, pp. 308-319, 2015.
[70] C. C. Lam, T. Alviar-Martin, S. A. Adler, and J. B. Y. Sim, "Curriculum integration in Singapore: Teachers' perspectives and practice," *Teach. Teach. Educ.*, vol. 31, pp. 23-34, April 2013.
[71] F. D. Davis, "Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology," *MIS Q.*, vol. 13, no. 3, pp. 319-340, Sep. 1989.
[72] G. Kabra, A. Ramesh, P. Akhtar, and M. K. Dash, "Understanding behavioural intention to use information technology: Insights from humanitarian practitioners," *Telemat. Informatics*, vol. 34, no. 7, pp. 1250-1261, 2017.
[73] S. Alharbi and S. Drew, "Using the technology acceptance model in understanding academics' behavioural intention to use learning management systems," *International Journal of Advanced Computer Science and Applications*, vol. 5, no. 1, pp. 143-155, 2014.
[74] M. A. Ramdhani, D. Sa'adillah Maylawati, A. S. Amin, and H. Aulawi, "Requirements Elicitation in Software Engineering," *Int. J. Eng. Technol.*, vol. 7, no. 2.29, pp. 772-775, 2018.
[75] S. Vijayarani, J. Ilamathi, and M. Nithya, "Preprocessing Techniques for Text Mining - An Overview," *Int. J. Comput. Sci. Commun. Networks*, vol. 5, no. 1, pp. 7-16, 2015.
[76] S. Kannan *et al.*, "Preprocessing Techniques for Text Mining," *Int. J. Comput. Sci. Commun. Networks*, vol. 5, no. 1, pp. 7-16, 2015.
[77] D. S. Maylawati, H. Aulawi, and M. A. Ramdhani, "Flexibility of Indonesian text pre-processing library," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 13, no. 1, pp. 420–426, 2019.
[78] H. A. Robbani, "Sastrawi," *MIT*, 2016.
[79] M. Akcayir, H. Dundar, and G. Akcayir, "What makes you a digital native? Is it enough to be born after 1980?," *Comput. Human Behav.*, vol. 60, pp. 435-440, 2016.

BIOGRAPHIES OF AUTHORS

Dian Sa'adillah Maylawati is a lecturer in the Department of Informatics at the UIN Sunan Gunung Djati Bandung, Indonesia. Her current research interests focus on Software Engineering, Expert System, Text Mining, and Natural Language Processing. She takes the Ph.D degree of Information and Communication Technology in Universiti Teknikal Malaysia Melaka (UTeM).
SCOPUS ID: 57200569961
ORCID ID: 0000-0002-1193-3370

Tedi Priatna is an associate professor at the Department of Islamic Education at the UIN Sunan Gunung Djati Bandung, Indonesia. His current research interests focus on Islamic Education.
SCOPUS ID: 57205019783
ORCID ID: 0000-0002-8491-5405

Hamdan Sugilar is a lecturer in the Department of Mathematics Education at the UIN Sunan Gunung Djati Bandung, Indonesia. His current research interests focus on Mathematics Education.
SCOPUS ID: 57200558086
ORCID ID: 0000-0001-8588-3372

Muhammad Ali Ramdhani is a Professor in Research of Information Technology in the Department of Informatics at the UIN Sunan Gunung Djati Bandung, Indonesia. His current research interests focus on Information System, Expert System, Decision Support System, Strategic Management, and Research Methodology.
SCOPUS ID: 54401502400
ORCID ID: 0000-0002-6492-067X