Groups Allocation Based on Sentiment-Epistemic Analysis in Online Learning Environment

by Hapnes Toba, Mewati Ayub, Maresha Caroline Wijanto , Roy Parsaoran, Ariyanto Sani

Groups Allocation Based on Sentiment-Epistemic Analysis in Online Learning Environment

Hapnes Toba
Faculty of Information Technology
Maranatha Christian University
Bandung, Indonesia
hapnestoba@it.maranatha.edu

Mewati Ayub
Faculty of Information Technology
Maranatha Christian University
Bandung, Indonesia
mewati.ayub@it.maranatha.edu

Maresha Caroline Wijanto
Faculty of Information Technology
Maranatha Christian University
Bandung, Indonesia
maresha.cw@it.maranatha.edu

Roy Parsaoran
Faculty of Information Technology
Maranatha Christian University
Bandung, Indonesia
1772044@maranatha.ac.id

Ariyanto Sani
Faculty of Information Technology
Maranatha Christian University
Bandung, Indonesia
1772004@maranatha.ac.id

***Abstract*—Collaborative learning methods in an online learning environment, encourages students to interact actively among themselves in a work group. Instructors or lecturers need to combine potential students to work together as a group, but this task is not easy since the characteristics of a student are sometimes not explicitly known. In this preliminary research, we propose a solution to answer this problem. Our methodology is composed in three steps. It begins with the sentiment analysis process with a textual history of online conversation or discussion. The next step is to classify the text into one of predefined epistemic groups. Further, we visualize the model in an epistemic network graph which is based on singular value decomposition. The group allocation is built based on *k*-means clustering. The case study in this paper is related to information technology-based subjects, and thus we classify our sentiment-epistemic analysis in three collaborative aspects, *i.e.*: project management, attitude and technology affinity. Our results show that by combining sentiment-epistemic analysis and *k*-means clustering, a holistic group allocation can be produced which would be beneficial in a collaborative learning environment.**

Keywords*—online learning, k-means, sentiment analysis, group allocation

I. INTRODUCTION

The main characteristic of a collaborative learning activity is to create a learning environment for students in a creative way inside a discussion group. Each student in the group is expected to develop his/her ability towards the learning objectives of a course. Interactions and discussions between students are the key to a successful collaborative learning approach [1]. In a small work group setting, students are encouraged to express their opinions, communicate their ideas and finally collaborate to solve alternatives of a problem given as their tasks. A student who has trouble understanding some aspects of the study materials may discuss them with other students in the group, and thus the whole group members will develop together [2][3].

Although it seems that collaborative learning is an ideal situation for student development, sometimes the situations in a workgroup are unbalanced. In reality, the possibility is very high that only one or two students are active, but the others are not really involved in group discussions [1]. Some students have difficulties in starting a conversation, but some are really talkative. In an online situation, for instance: in a forum discussion or chatting via instant messenger, the situation could be better in the presence of an instructor or lecturer whose job is to observe the discussion and suggests study resources [2]. Online discussion settings might also enhance students' performance, since the directed discussions in the group lead to the positive implication in student engagement [3].

Group members' allocation is an important step to ensure a successful collaborative learning activity. Students might choose their peers independently, but this strategy has a main drawback, i.e. exclusivity, not every student has the willingness to freely include other students outside their close friends. Instructors or lecturers need to combine potential students to work together in a group, but this task is not easy since the characteristics of a student are sometimes not explicitly known.

In this research, we propose an initial solution to answer this problem. Our approach is based on textual features which analyzed further into sentiment [4] and epistemic classification [5]. An important resource to reveal students' characteristics is their historical interactions in the course learning system (CLS) [6], especially in forum discussions or other online chat activities, such as in social media or instant messenger.

II. RELATED WORK

Sentiment analysis or opinion mining is a text classification approach which analyzes people's subjectivity or emotional opinion in a particular situation based on textual features, such as in product reviews [4]. The common way to express sentiments is in a positive, neutral or negative expression. In special cases, sentiment analysis could also be expanded to analyze expressions in other categories, such as in educational situations [7]. Popular techniques in sentiment analysis are employing machine learning algorithms, such as: Naive Bayes, decision trees and deep learning approaches.

By performing sentiment analysis, we also could analyze typical words used by someone to express their feelings and how they react to comments from others [8]. In a group discussion situation, students' performance could be analyzed by using the expressions they presented in conversations. Some students are active in question-answer activities, while the others might be active to invite their peers in completing the tasks or assignments. This is exactly what is offered by epistemic network analysis (ENA).

ENA is a method to identify, measure and visualize the connection between elements in a textual collection. The fundamental ingredient in ENA is a text classification which further is visualized by using the weights of each component in terms of eigenvalues after the dimensionality reduction process by using latent semantic analysis [9]. The classification concept in ENA is driven by the epistemic category of each word. For instance, the word 'Windows' tends to show technology aspects in IT-related subjects, or the word 'tomorrow' indicates how someone is aware to fulfil the schedule of an assignment. However, ENA needs to be enhanced so that we can see the actual influential words in each text category. In that sense, we might employ clustering methods, such as *k*-means, to dive deeper into the influential words in each epistemic category [10].

In this research a method is proposed which employs sentiment analysis technique to classify epistemic categories. These sentiments-epistemic classifications will be further analyzed to form a number of clusters which suggest student group membership in a collaborative learning situation. Unlike popular approaches in forming students' work groups which employed genetic algorithms with a large dataset from Massive Open Online Course (MOOC) [11][12], our approach is intended to work with small or large dataset based on the textual features.

III. Research Method

The overall research method can be seen in Fig. 1. The dataset is first pre-processed and tagged according to sentiment: positive or negative. After that a sentiment analysis model is built which will be used to analyze a new input of text. The epistemic network analysis is used to plot how the textual input in the dataset is emphasized. Finally, by employing *k*-means clustering a work group allocation is suggested. There are three items that need to be analyzed in the dataset: the sentiment (polarity) and epistemic category of each chat history, the overall tendency of the whole dataset whether it has more emphasis on one of the epistemic categorization, and the clusters of students who has produced the chat dataset.

A. Sentiment Analysis

The dataset consists of a list of textual chat history from a discussion forum or chat group. The list is composed by the students' name and the content of their responses. After the dataset is collected, two persons manually annotate the textual contents into sentiment analysis polarities: positive and negative categories. Besides the annotation of the polarities, epistemic categories are also tagged each chat or discussion entry in the dataset. Since the case study is related to information technology-based subjects, we classify our sentiment-epistemic analysis in three collaborative aspects, *i.e.*: project management, attitude and technology affinity [2][3][6]. These collaborative aspects are flexible and can be decided based on the research focus agreement.

Before the classification model of the sentiment-epistemic is built, a pre-processing activity needs to be accomplished. The pre-processing process consists of the following sub-activities: remove stop-words, remove slang words and abbreviations, and stemming.

B. Random Forest

The sentiment analysis model is built by using the random forest algorithm [13]. Random forest is one of the finest algorithms in text classification since it has the ability to choose the best decision tree which consists of important words in an ensemble style [14][15].

C. Epistemic Network Analysis

The results of the sentiment analysis will be analyzed further by the ENA to see how the tendency of the whole dataset, whether it has more emphasis on one of the epistemic categorizations, i.e.: project management, attitude and technology affinity in our case. The basic feature of ENA is to compute how the words in each epistemic category related to each other in the dataset [5][9], but not to analyze the relation among the chat categories itself.

D. K-means

To analyze how the students are really connected in each sentiment-epistemic category, we need further analysis methods. For this purpose, we employ the *k*-means clustering method. With the *k*-means method, we can create several student's clusters from sentiment-epistemic frequency matrices. The rows in this matrix consist of the students' name, and the columns are the number of times each student is categorized in a specific sentiment-epistemic category from their chat history in the dataset. Based on the *k*-means results, the work groups for all students can be allocated.

IV. Experiment, Result and Discussion

In this section details of the experiments explained in Fig. 1 is described. Related results from each part of the experiment will also be discussed.

A. Dataset

The dataset for the experiments is taken from two sources, i.e.: forum discussion from our faculty CLS and WhatsApp Group (WAG). There are three under-graduate subjects for the even semester 2019/20 taken in the datasets, i.e.: Enterprise Resource Planning (ERP), Project Management (PM) and Research Methodology (RM). The general statistics of the dataset are given in Table I. This dataset is designed by involving selected students, which follow all the courses and by manual observation have given adequate-to-excellent individual and group performances. The students are chosen from arbitrary work groups. By design, we do not want to know whether they are coming from the same group or not. It is expected that by observing these students we will have a balanced students' group distribution with special characteristics in the clusters which is formed during the *k*-means step.

TABLE I. GENERAL STATISTICS OF THE EXPERIMENT DATASET

Subjects	#Students Involved	#Chats	#Words Total	Source
ERP	5	293	901	WAG
PM	5	388	1179	WAG & CLS
RM	3	148	987	WAG

Two annotators are assigned to label the chat history dataset into two aspects, i.e.: positive and negative polarity; and also, the epistemic categories: project management, attitudes, and technology affinity. The Kappa Cohen inter-annotator agreement during tagging is measured, with the value 0.63, which can be interpreted as substantial-moderate [16]. Snapshot examples of the annotation results of three students' chats can be seen in Table II.

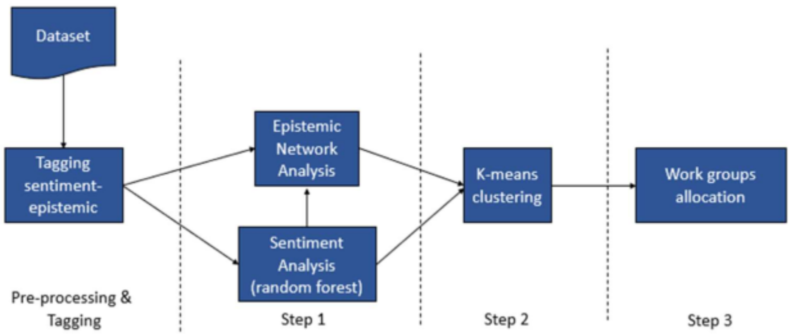

Fig. 1. The main steps in the research.

TABLE II. EXAMPLES OF ANNOTATION RESULTS: EACH CHAT HISTORY CAN BE CATEGORIZED INTO MORE THAN ONE EPISTEMIC CATEGORIES

Students	Chats (translated into English)	Tagged as (sentiment and epistemic)
A	Just open the drive and share it via email... then create a word file in that drive.	Polarity: + (identified by the word: "create") Epistemic: technology (identified by the word: "file")
B	I still cannot understand... Do we need to finish the assignment in just three days?	Polarity: - (identified by the word: "cannot") Epistemic: project management (identified by the word: "assignment")
C	Keep the spirit up, guys! It's almost done.	Polarity: + (identified by the word: "keep up") Epistemic: attitude (identified by the word: "spirit")

B. *Sentiment and Epistemic Networks Analysis*

The sentiment-epistemic model is built by using a random forest algorithm with unlimited depth and 5-cross validation (CV). The mean accuracy of the 5-CV is 60.76%. One of the main challenges during the model creation is the multilabel nature of the epistemic categories. Another issue is that some technology jargons are not uniformly written, for instance the terms Google Drive, sometimes is written as: gdrive, G-Drive, or even just written as a drive.

The epistemic visualization of the sentiment-epistemic annotation can be seen in Fig. 2. We visualize only the dataset for the ERP and PM subjects. The data come from the WAG chats, so it is expected that the visualization reveals the characteristics of short messages in the subjects. The software development process of the system in our CLS is published in [17].

In Fig. 2., the direction of the terms in the ERP chats history (red lines) is dominated with the 'positive-attitudes', but surprisingly it shows the negative technological affinity. This fact might suggest that the students need more guidance to complete their assignments during the practical works. The blue lines indicate how the students interact during the PM sessions. In line with the subject, the students have shown their ability to manage their assignments, since the direction of the terms is dominated with the 'positive-project management'.

C. *Students' k-means Clusters*

To produce the *k*-means clusters, a matrix which consists of sentiment-epistemic frequencies for each student is created. The matrix can be seen in Table III. From this matrix, the *k*-means algorithm suggests k=2 as the optimal number of clusters (see Fig. 3), by employing the elbow method cluster evaluation [18]. However, to evaluate the distribution of students in the clusters, we also run k=3 and k=4. The results of the clusters can be seen in Fig. 4.

Student clusters (before allocation) are described as follows:

- k=2 (Fig. 4 (a)):
  - Cluster 1 (red) = B, C, **I, J, K, L, M**
  - Cluster 2 (blue) = **A, D, E**, F, G, H
- k=3 (Fig. 4 (b)):
  - Cluster 1 (red) = F, G, H
  - Cluster 2 (green) = **A, D, E**
  - Cluster 3 (blue) = B, C, **I, J, K, L, M**
- k=4 (Fig. 4 (c)):
  - Cluster 1 (red) = **A, D, E**
  - Cluster 2 (green) = B, C, G, H
  - Cluster 3 (blue) = F
  - Cluster 4 (purple) = **I, J, K, L, M**

Students in the bold are staying in the same clusters after the allocation.

From the results in Fig. 4, we can observe that there are several students, which stay in the same clusters for all values of k's. Those students are: (A, D, E) and (I, J, K, L, M). Other students are rather dynamic and not staying in the same clusters, they are: (B, C, F, G, H). These results suggest that the students who stay in the same clusters are close one to the other and have the same characteristics. The students who dynamically changed their membership during the clusters' forming are those who have the tendency for flexible grouping.

TABLE III. THE SENTIMENT-EPISTEMIC FREQUENCIES MATRIX DURING THE *K*-MEANS PROCESS

	Pro. Man.		Attitudes		Technology	
	+	-	+	-	+	-
A	1	1	48	10	24	5
B	1	2	37	4	8	2
C	1	3	20	5	3	2
D	6	1	71	15	24	6
E	4	0	66	16	20	0
F	31	5	53	7	34	2
G	17	4	13	11	9	5
H	13	2	29	7	17	1
I	3	0	2	0	0	0
J	0	0	2	0	1	0
K	0	0	1	0	0	0
L	0	0	1	0	0	0
M	0	0	5	0	0	0

Further analysis based on Table III, we can see that the students in the set (A, D, E) have strong characteristics in attitudes and technical affinity. The student set (I, J, K, L, M) contains passive students in the chat history, they only commented on important issues. Based on this observation, we may summarize that the students which are clustered together, for any value of k, need to be separated for work group allocation. In this way we will have a mixed composition of students with various characteristics which would be beneficial in a collaborative learning environment.

D. Students Workgroups Allocation

A basic algorithm to allocate work group members for any number of k (i.e. the minimum number of a group membership), the following allocation algorithm is used:

```
Algorithm AllocateStudentWorkGroup
Input:
clusters c, a member of k-means clusters
Process:
prepare w, the number of work groups (i.e.
the same as k)
for all clusters in c1...ck
   extract all students s member of a
   cluster ck
   for all work group w in w1...wk
      for all students s in s1...si in ck
          assign si consecutively into
          w1...wk
Output:
return w
```

In this basic algorithm all students are distributed evenly and consecutively in all workgroups. It is expected that in this basic algorithm we will have a combination of mixed students' characteristics. For our example in Fig. 4 (c), for the number of $k = 3$, we will have for instance the following work groups: Group 1 consists of students: A, B, F, J, M; Group 2: C, D, G, K; and finally Group 3: E, H, I, L.

This allocation is validated with the actual group membership during the course. By using the percentage of coverage evaluation, we have observed that: Group 1 has (A, B and M, *i.e.* 60% coverage) working together as a group, Group 2 has (C, D and K, *i.e.* 75% coverage), and Group 3 has (E, H, and L, *i.e.* 75% coverage). These initial results suggest that by combining sentiment and epistemic analysis with the k-means clustering strategy, we might have appropriate students' work group allocation. Further research needs to be done to validate the students' group assignment, for instance by calculating the sentiment-epistemic distances between the students, and so we might produce a relational graph between the students.

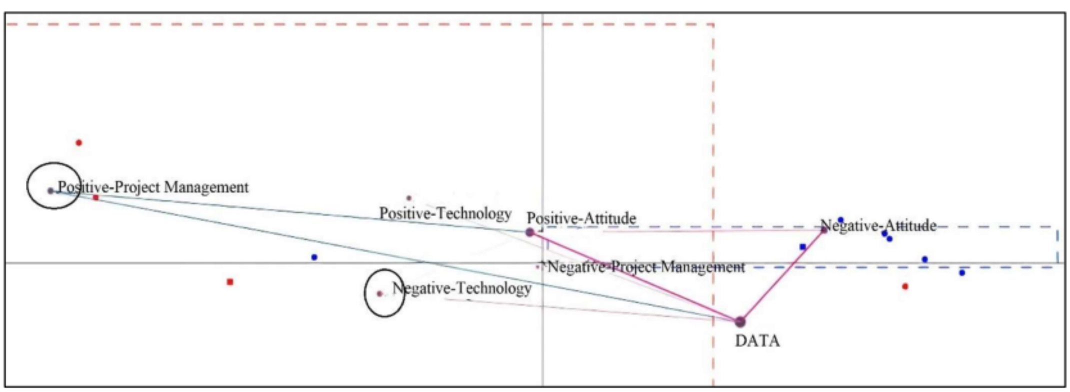

Fig. 2. The visualization of ENA from the ERP and PM subjects. In the figure we can see the red lines are the epistemic directions of the ERP chats history, and the blue lines are the directions of the PM chats history

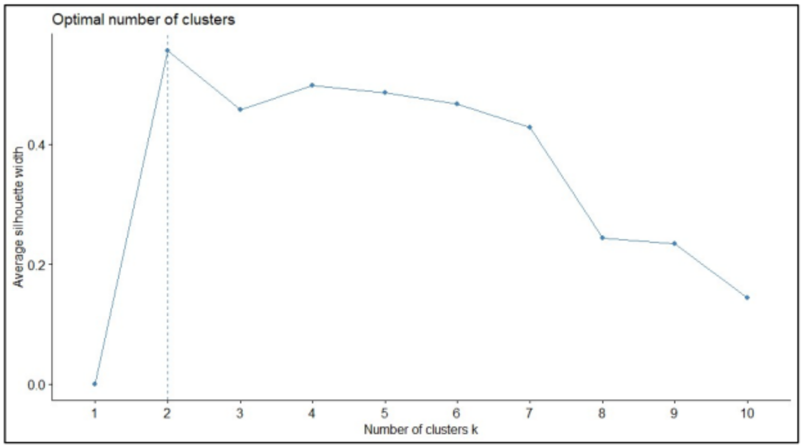

Fig. 3. Number of optimal clusters (k=2) according to the dataset used in the experiments

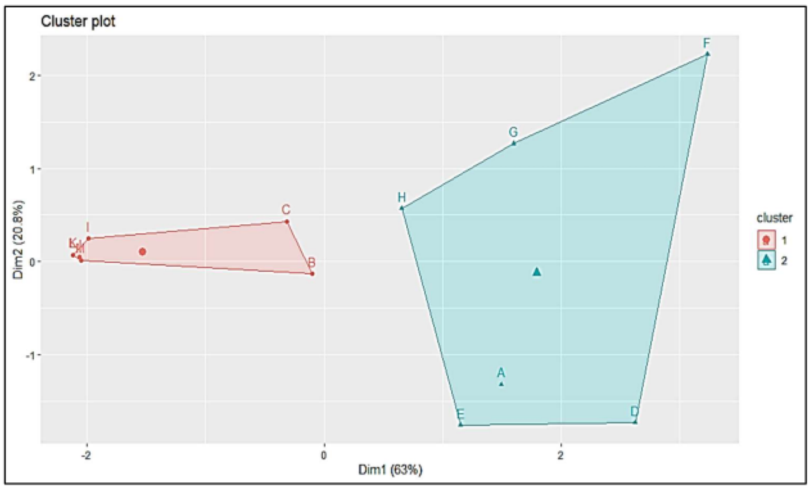

Fig. 4 (a). Results of k-means students' clustering with k=2

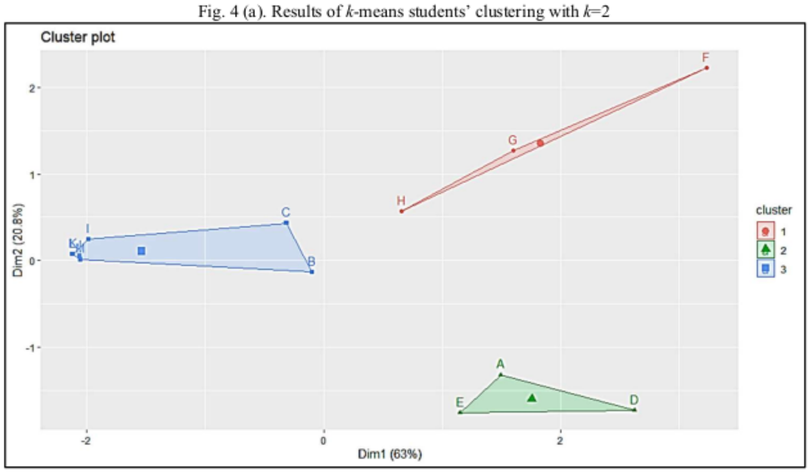

Fig. 4 (b). Results of k-means students' clustering with k=3

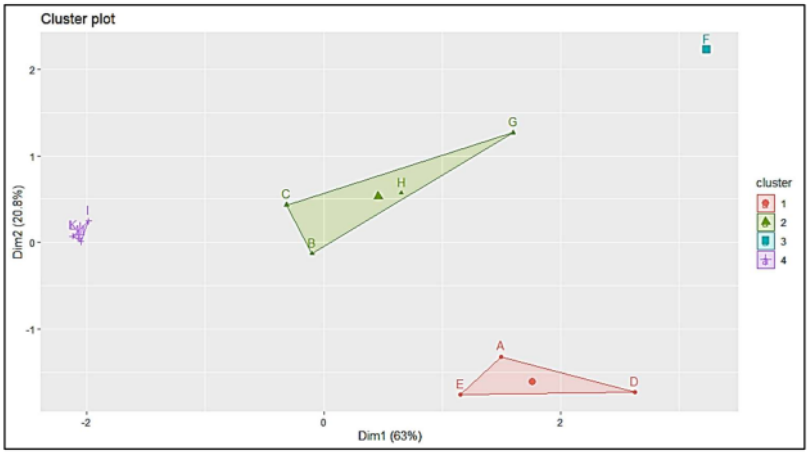

Fig. 4 (c). Results of *k*-means students' clustering with *k*=4

V. Conclusion and Future Work

In this research, we demonstrate that by combining sentiment and epistemic analysis of textual chat history with the *k*-means clustering strategy, we might have appropriate students' work group allocation. In this way, lecturers or instructors will have a mixed composition of allocated students' with various characteristics and would be beneficial in a collaborative learning environment.

Further research needs to be carried out to validate the students' group assignment, for instance by forming relational graphs between the students. In particular, our approach has shown some opportunities to enhance natural language processing techniques for sentiment analysis in educational situations, for instance by performing aspect-based models.

Acknowledgment

This research is supported by Ministry of Education, Culture, Research and Technology Republic of Indonesia, Grant Number 1868/E4/AK.04/2021, July 12th, 2021; 005/SP2H/RT-MONO/LL4/2021, July 19th, 2021; 235-D/LPPM/UKM/VII/2021, July 21st, 2021.

References

[1] H. Jeong and C. E. Hmelo-Silver, "Seven Affordances of Computer-Supported Collaborative Learning: How to Support Collaborative Learning? How Can Technologies Help?," *Educational Psychologist*, vol. 51, no. 2, pp. 247–265, 2016.

[2] B. D. Wijanarko, D. F. Murad, Y. Heryadi, Lukas, H. Toba, and W. Budiharto, "Questions Classification in Online Discussion Towards Smart Learning Management System," in *2018 Proc. International Conference on Information Management and Technology (ICIMTech)*, 2018, pp. 251-255, doi: 10.1109/ICIMTech.2018.8528131.

[3] M. Ayub, H. Toba, M. C. Wijanto, and S. Yong, "Modelling online assessment in management subjects through educational data mining," in *2017 Proc. International Conference on Data and Software Engineering (ICoDSE)*, 2017, pp. 1-6, doi: 10.1109/ICODSE.2017.8285881.

[4] B. Liu, *Sentiment analysis and opinion mining*. San Rafael, Calif. (1537 Fourth Street, San Rafael, CA 94901 USA): Morgan & Claypool, 2012.

[5] R.F. Mello and D. Gašević, "What is the effect of a dominant code in an epistemic network analysis?", in *2019 Proc. International Conference on Quantitative Ethnography (ICQE)*, 2019, pp. 66-76, doi: https://doi.org/10.1007/978-3-030-33232-7_6.

[6] M. Ayub, H. Toba, S. Yong, and M.C. Wijanto, "Modelling students' activities in programming subjects through educational data mining", *Global Journal of Engineering Education*, vol. 19, no. 3, pp. 249-255, 2017.

[7] K. Mite-Baidal, C. Delgado-Vera, E. Solís-Avilés, A. Herrera Espinoza, J. Ortiz-Zambrano, and E. Varela-Tapia, "Sentiment analysis in education domain: A systematic literature review," in *2018 International Conference on Technologies and Innovation (CITI)*, 2018, pp. 285-297, doi: 10.1007/978-3-030-00940-3_21.

[8] V. R. Attili, S. R. Annaluri, S. R. Gali, and R. Somula, "Behaviour and Emotions of Working Professionals Towards Online Learning Systems," *International Journal of Gaming and Computer-Mediated Simulations*, vol. 12, no. 2, pp. 26–43, 2020.

[9] D. W. Shaffer, W. Collier, and A. R. Ruis, "A Tutorial on Epistemic Network Analysis: Analyzing the Structure of Connections in Cognitive, Social, and Interaction Data," *Journal of Learning Analytics*, vol. 3, no. 3, pp. 9–45, 2016.

[10] J. Yang and J. Wang, "Tag clustering algorithm LMMSK: improved K-means algorithm based on latent semantic analysis," *Journal of Systems Engineering and Electronics*, vol. 28, no. 2, p. 374, 2017.

[11] A. Sukstrienwong, "A Genetic-algorithm Approach for Balancing Learning Styles and Academic Attributes in Heterogeneous Grouping of Students," *International Journal of Emerging Technologies in Learning*, vol. 12, no. 3, p. 4-25, 2017.

[12] T. K. Shih, W. K. T. M. Gunarathne, A. Ochirbat, and H.-M. Su, "Grouping Peers Based on Complementary Degree and Social Relationship using Genetic Algorithm," *ACM Transactions on Internet Technology*, vol. 19, no. 1, pp. 1–29, 2019.

[13] L. Breiman, "Random forests", *Machine learning*, vol. 45 no. 1, pp.5-32, 2001.

[14] T. Pranckevičius and V. Marcinkevičius, "Comparison of Naive Bayes, Random Forest, Decision Tree, Support Vector Machines, and Logistic Regression Classifiers for Text Reviews Classification," *Baltic Journal of Modern Computing*, vol. 5, no. 2, pp. 221-232, 2017.

[15] A. Onan, S. Korukoğlu, and H. Bulut, "Ensemble of keyword extraction methods and classifiers in text classification," *Expert Systems with Applications*, vol. 57, pp. 232–247, 2016.

[16] C. Grisot, "A quantitative approach to conceptual, procedural and pragmatic meaning: Evidence from inter-annotator agreement," *Journal of Pragmatics*, vol. 117, pp. 245–263, 2017.

[17] R. Parsaoran, R. and H. Toba, "Rekomendasi Pembentukan Kelompok Kerja bagi Mahasiswa berbasis Sentimen-Epistemic Analysis [Work groups recommendation for students based on Sentiment-Epistemic Analysis]". *Jurnal STRATEGI*, vol. 3, no. 1, pp.109-121, 2021.

[18] C. Yuan and H. Yang, "Research on K-Value Selection Method of K-Means Clustering Algorithm," *J*, vol. 2, no. 2, pp. 226–235, 2019.

Groups Allocation Based on Sentiment-Epistemic Analysis in Online Learning Environment

ORIGINALITY REPORT

8%	5%	6%	4%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.icicel.org Internet Source	1%
2	Zackary P. T. Sin, Ye Jia, Astin C. H. Wu, Isaac Dan Zhao et al. "Towards an Edu-Metaverse of Knowledge: Immersive Exploration of University Courses", IEEE Transactions on Learning Technologies, 2023 Publication	1%
3	journal.iis.sinica.edu.tw Internet Source	1%
4	Submitted to University of York Student Paper	1%
5	comptes-rendus.academie-sciences.fr Internet Source	1%
6	Ramy Elmoazen, Mohammed Saqr, Matti Tedre, Laura Hirsto. "A Systematic Literature Review of Empirical Research on Epistemic Network Analysis in Education", IEEE Access, 2022 Publication	<1%

7	Wentao Fan, Fuyuan Xiao. "A complex Jensen–Shannon divergence in complex evidence theory with its application in multi-source information fusion", Engineering Applications of Artificial Intelligence, 2022 Publication	<1%
8	www.researchgate.net Internet Source	<1%
9	Dragan Gašević, Srećko Joksimović, Brendan R. Eagan, David Williamson Shaffer. "SENS: Network analytics to combine social and cognitive perspectives of collaborative learning", Computers in Human Behavior, 2018 Publication	<1%
10	Malak Chniter, Abir Abid, Ilhem Kallel. "Towards a Bio-inspired ACO Approach for Building Collaborative Learning Teams", 2018 17th International Conference on Information Technology Based Higher Education and Training (ITHET), 2018 Publication	<1%
11	Nafis Irtiza Tripto, Mohammed Eunus Ali. "Detecting Multilabel Sentiment and Emotions from Bangla YouTube Comments", 2018 International Conference on Bangla Speech and Language Processing (ICBSLP), 2018 Publication	<1%

12	Simanta Hazra, Sunil Karforma, Abhishek Bandyopadhyay, Sayantani Chakraborty, Debasis Chakraborty. "Prediction of Crop Yield Using Machine Learning Approaches for Agricultural Data", Institute of Electrical and Electronics Engineers (IEEE), 2023 Publication	<1 %
13	Xiaohui Tao, Aaron Shannon-Honson, Patrick Delaney, Lin Li, Christopher Dann, Yan Li, Haoran Xie. "Data Analytics on Online Student Engagement Data for Academic Performance Modeling", IEEE Access, 2022 Publication	<1 %
14	e-tarjome.com Internet Source	<1 %
15	ijcjournal.org Internet Source	<1 %
16	munin.uit.no Internet Source	<1 %
17	www.monash.edu Internet Source	<1 %
18	Abiodun M. Ikotun, Absalom E. Ezugwu, Laith Abualigah, Belal Abuhaija, Jia Heming. "K-means Clustering Algorithms: A Comprehensive Review, Variants Analysis, and	<1 %

Advances in the Era of Big Data", Information Sciences, 2022

Publication

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	On		

Groups Allocation Based on Sentiment-Epistemic Analysis in Online Learning Environment

GRADEMARK REPORT

FINAL GRADE

/0

GENERAL COMMENTS

Instructor

PAGE 1

PAGE 2

PAGE 3

PAGE 4

PAGE 5

PAGE 6