Sentiment Analysis Using the Support Vector Machine For Community Compliance Representation in The Covid-19 Pandemic Period

Eneng Tita Tosida, Erniyati, Krisna
Department of Computer Science, Universitas Pakuan
Jl. Pakuan PO Box 452, 16143 Bogor, Jawa Barat, Indonesia
enengtitatosida@unpak.ac.id; neni_erniyati@yahoo.com; tkjkrisna@gmail.com

Abdul Talib Bon
Department of Production and Operations, University Tun Hussein Onn Malaysia, Malaysia
talibon@gmail.com

Abstract

The increase in positive cases of Covid-19 in Indonesia is still relatively high. The main factor affecting the high number of positive cases of Covid-19 in Indonesia is the decreasing level of citizen compliance. This study aims to build a model of citizen compliance representation through a sentiment analysis approach using social media. The method used to build this model is done with a text mining approach, using the sentiment analysis by Support Vector Machine (SVM) algorithm. This model was built using the term Large-Scale Social Restrictions (LSSR) as a key phrase. This key phrase is interpreted through four main types of activities, namely work from home (WFH), studying at home, transportation, and physical distancing. We used data from twitter and news period April-August 2020. This representative model of citizen compliance during the Covid-19 pandemic is able to show a visualization of the condition of citizens' compliance levels grouped by city and province. The citizen compliance representation model shows an accuracy rate of 98% for WFH activities, 93% for learning activities at home, 88% for transportation activities and 90% for physical distancing activities. This model can be an initial reference for identifying the level of compliance of citizens with regard to government policies related to programs for handling Covid-19 cases in Indonesia.

Keywords:
Citizen compliance, Covid-19, Representative model, Support Vector Machine, Sentiment analysis

1. Introduction

Currently, the world is experiencing epidemic that hitsalmost all countries. The disease outbreak is Covid-19 or commonly known as the corona virus. Corona viruses are a large family of viruses that can cause disease in animals or human. In human, the corona virus is known to cause respiratory infections ranging from the common cold to more severe diseases such as Middle East Respiratory Syndrome (MERS) and Severe Acute Respiratory Syndrome (SARS) (WHO South-East Asia, 2020). Every day the data on positive cases and deaths in Indonesia is increasing. Various efforts have been made by the government to reduce the rate of transmission and reduce wider spread of the Covid-19 virus to minimize positive victims and deaths in various regions in Indonesia. The efforts that have been implemented to date are Physical Distancing and Large-Scale Social Restrictions (LSSR). However, there are still many people who have not comply with the government regulation. This research aims to implement Sentiment Analysis using a Support Vector Machine for Representation of Public Compliance during the Covid-19 Pandemic. Sentiment analysis or sentiment on mining or sentiment on extraction is a field of study that analyzes opinions, sentiments, evaluations, attitudes and emotions towards an entity such as products, services, organizations, individuals, problems, topics and attributes of that entity. Furthermore, sentiment analysis can express emotional sadness, joy, or anger (Liu, 2012). In this study, we used data from social media Twitter with the keywords "obey

LSSR study at home", "obey LSSR WFH", "comply with LSSR transportation" and "comply with physical distancing".

The data is processed using text mining and SVM classification. Text mining is a technology used to analyze unstructured data in the form of text. In text mining analysis, there are two main phases, namely preprocessing and integration of unstructured data, the second is statistical analysis of data that has been preprocessed to extract content from what is contained in the text (Francis and Flynn, 2010). The level of accuracy in the model generated by the transition process with SVM highly depends on the kernel function and the parameters used (Han, Kamber 2006). Text mining is a part of data mining where the process is mainly carried out by extracting knowledge and information from patterns contained in a set of text documents using certain analytical tools (Monarizqa etc., 2014).

Preprocessing is the stage that is carried out before the classification stage. Previously, the raw dataset was cleaned. This stage is carried out to simplify the classification process (Apasari, 2017). Then the word weighting is carried out using the TF-IDF method, follow by Support Vector Machine (SVM) classification which is a learning system using space in the form of linear functions in a high-dimensional feature space that is trained using learning algorithms based on optimization theory by implementing learning bias.

SVM has the highest level of accuracy in terms of text classification. In a journal entitled sentiment analysis of Twitter data: a survey of techniques; the results of the classification method with the highest accuracy level are using the SVM with an accuracy of 76.68%, Maximum Entropy of 74.93% and Naïve Bayes Classifier (NBC) of 74, 65% (Kharde and Sonawane, 2016). SVM is a method in machine learning that works with the principle of Structural Risk Minimization (SRM) with the aim of finding the best hyperplane that separates two classes in the input space. The main principle of using SVM is to find the best hyperline that separates two classes in the input space. The hyperline can be a line in two dimensions and can be a flat plane in multiple planes.The input in the SVM is a vector data consisting of real numbers. Meanwhile, each label is denoted by $yi \in \{-1, +1\}$ with i = 1,2,3..l, where 2 l is the amount of data. After that, a validation test is performed using confusion matrix. The confusion matrix is a useful tool for analyzing how well the classifier recognizes tuples from different classes (Han and Kamber, 2006). The objectives of this study are to determine the level of compliance of the Indonesian people with government regulations in handling the spread and transmission of Covid-19, classifying sentiment analysis using SVM which represent public compliance during the Covid-19 pandemic and to find out the accuracy of the method.

2. Research Methods

2.1 The object being analyzed

Tweet data and data from news portals with regard to "obey LSSR study at home", "obey LSSR WFH", "comply with LSSR transportation" and "comply with physical distancing".

2.2 Research steps

The research steps can be described as follow:

Step 1 : Tweet Data Collection

Tweet data is obtained by using the capture to crawl the data in the tweet. The data search was based on the query that was inputted, namely regarding LSSR compliance which included "complying with LSSR studying at home", "complying with LSSR WFH", "obeying LSSR for transportation" (Carteni et al. 2020) and "obeying physical distancing".

Step 2 : Data Cleaning and Integration.

Data cleaning is a process of identifying incomplete or irrelevant information and then modifying or deleting dirty data (Saini, 2019). Before classifying text documents, a preprocessing stage is necessary. The Twitter data and news obtained are not fully ready to be used for the direct classification process because the data is still not well-structured and there is a lot of noise. The data still contains numbers, punctuation marks, emoticons, and other words that are less meaningful to be used as features. Therefore, it is necessary to do preprocessing which aims to construct uniform words, eliminate characters other than letters, and reduce the volume of vocabulary so that the data will be more structured.

Step 3 : Data Selection and Data Transformation.

Data selection and transformation aims to find the features stored in important data based on the required requirements. This process also aims to reduce the number of variables and data that are not really needed and reformat the data or combined into a format suitable for processing in data mining.

Step 4 : Pattern Evaluation.

It evaluates the exploited pattern to interpret and extract knowledge from the sought pattern by visualizing the pattern (Francesco Gullo, 2015). Pattern Evaluation is carried out to identify interesting patterns that exist in the data

into the found knowledge based. In this stage, the results of the data mining techniques are in the form of distinctive patterns and sentiment analysis models that are evaluated to assess whether the existing hypothesis has indeed been reached.

Step 5 : Presentation of Knowledge.

Presentation of knowledge that has been generated is poured in the form of visualization to make it easier for users to read the results of the hypothesis.

2.3 Support Vector Machine

Support Vector Machine (SVM) is proposed as an alternative to standard SVM which has proven to be more efficient than traditional SVM in large-scale data processing (Huang et al., 2014). Support Vector Machine (SVM) is a classification technique in data mining. SVM is a method in machine learning that works with the principle of Structural Risk Minimization (SRM) with the aim of finding the best hyperplane that separates two classes in the input space. The SVM has other benefits such as the model being built has an explicit dependence on a subset of datapoints, as well as support vectors that can help in model interpretation. The main principle of SVM is to find the best hyperline that separates two classes in the input space. The hyperline can be a line in two dimensions and can be a flat plane in multiple planes. SVM is a machine learning tool that conducts training using training datasets, generalizes and makes predictions from new data. The steps in the SVM method are as follow:

1. define data points : $X_i = (X_1 + X_2 + \cdots + X_n) \in R_n$ (1)
2. define the data class : $y_i \in \{-1, +1\}$ (2)
3. data pairs and class : $\{(x_i, y_i)\}_{i=1}^{N}$ (3)
4. maximize function : $Ld = \sum_{i=1}^{N} a_i - \sum_{i=1}^{n} \sum_{i=1}^{n} a_i a_j y_i y_j K(x_i\, y_i)$ (4)

   conditions : $0 \leq a_i \leq C \; and \; \sum_{i=1}^{n} a_i y_i = 0$ (5)
5. count the value w dan b : w= $\sum_{i=0}^{n} a_i y_i x_i \quad b = -\frac{1}{2}(w.x^+ + w.x^-)$ (6)
6. classification decision function : $f(x) = w.x + b \; or \; f(x) = \sum_{i=0}^{n} a_i y_i K(x, x_i) + b$ (7)

3. Results and Discussion

3.1 Classification Results

After taking data from social media Twitter which was inputted according to the data retrieval keywords, including "obeying physical distancing", "obeying LSSR study at home", "obeying LSSR WFH" and "obeying LSSR transportation", the total data of 3797 was obtained. Data analysis results in the following classifications :

3.1.1 SVM Classification LSSR Study at Home

Based on the results of the sentiment analysis towards complying with the LSSR study at home, indicates that the obedience sentiment and disobedience were 880 and 524 comments from the overall result which can be seen in Fig.1. Apart from the overall results, the classification is also presented by city in Fig. 2 and province in Fig. 3. Tosida et al. (2019) developed learning models and media based on AR and integrated with animated films (Tosida et al. 2020b; Tosida et al. 2020c) that are able to support student learning so that it is more attractive and enjoyable. This model fits the atmosphere of the pandemic (Tosida et al. (2020a).

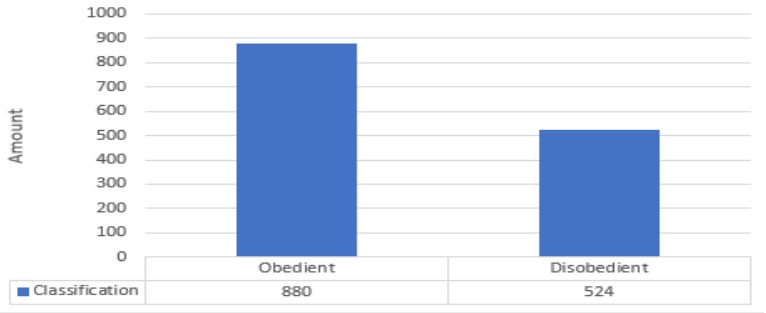


Figure 1. Classification Result of the LSSR Study at Home Overall

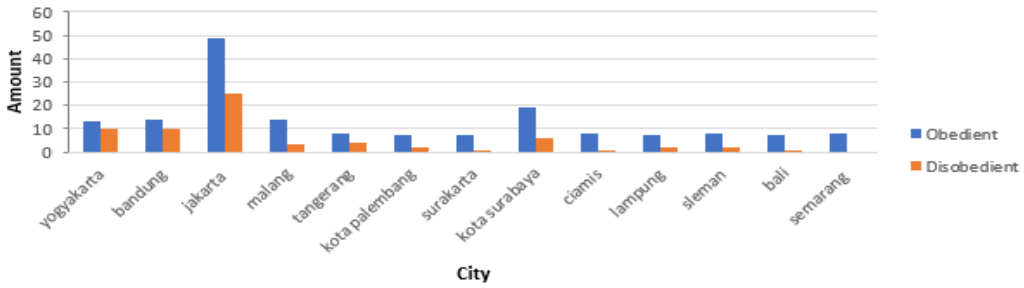


Figure 2. Classification Result of LSSR Study at home by city

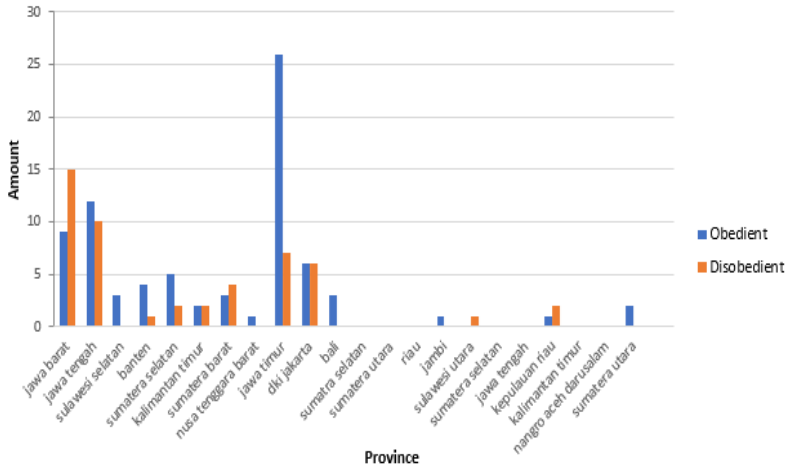


Figure 3. Classification Results of the LSSR Study at Home Base on Province

From these results, it turns out that the average community activities follow the government's regulation for working at home. Although in several cities and provinces, community compliant to the health protocol is below expectation. For example, the city of Jakarta as seen in Fig.2, the community compliant is above average.

3.1.2 WFH LSSR Sentiment Classification

The of the sentiment analysis towards WFH compliance showed 326 and 135 comments of obedience and non-compliant sentiments from the overall results which can be seen in Fig. 4.Considering the social and economic conditions, the existence of this WFH regulation will increase Indonesia's average poverty rate at the end of 2020. It means approximately 8 million people to experience new poverty due to this outbreak (Suryahadi, 2020). Apart from the overall results, the classification results are also presented by city in Fig. 5 and province in Fig. 6. From these results, it turns out that most of the community members follow the government regulation for working at home, although there are several cities and provinces that still do not comply with these regulations. For example, in Fig. 5 can be seen that the city of Jakarta is the most compliance city in Indonesia with regard to government health protocol regulations.

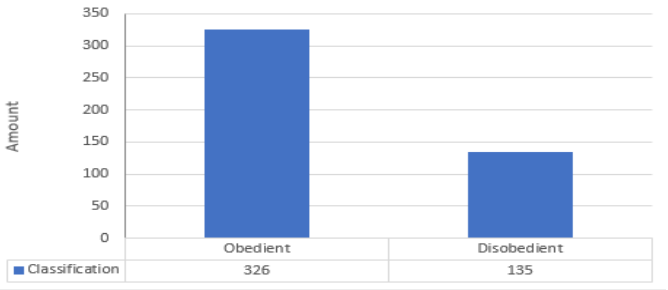


Figure 4. Classification Result of WFH LSSR Overall

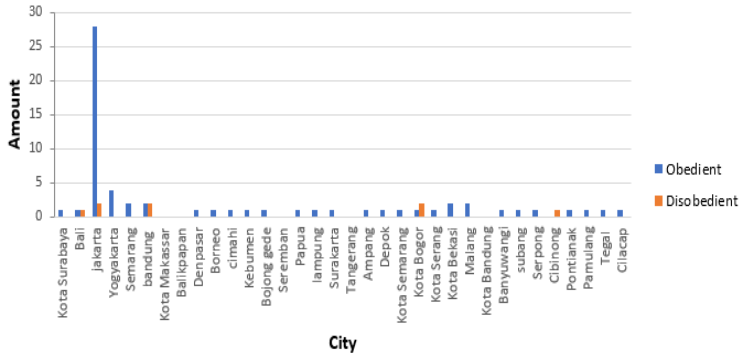


Figure 5. Classification Result of LSSR WFH by City

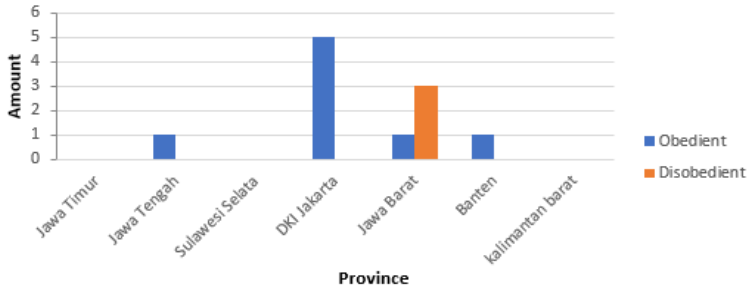
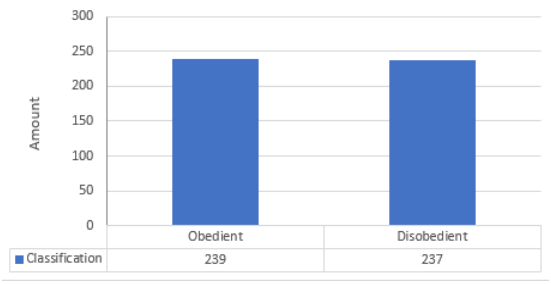


Figure 6. Classification Result of LSSR WFH Base on Province

3.1.3 LSSR Sentiment Classification for Transportation

LSSR limits the local mobility of the communities based on locality certain duration of timeframe. LSSR recommends also not to travel outside the residential area or returning from the hometown. The LSSR prohibition is implemented for residents in areas such as Jabodetabek, Bandung Raya area, Makassar, Pekanbaru, Tegal, Banjarmasin, Tarakan, Surabaya, Gowa Regency, Sidoarjo Regency, Gresik Regency and West Sumatra Province (Permana, 2020). Based on the results of the sentiment analysis towards transportation compliant, there are 239 and 237 comments of obedient and disobedient sentiments from the overall data which can be seen in Fig.7. Apart from the overall results, the classification results are also presented by city and province as in Fig. 8.



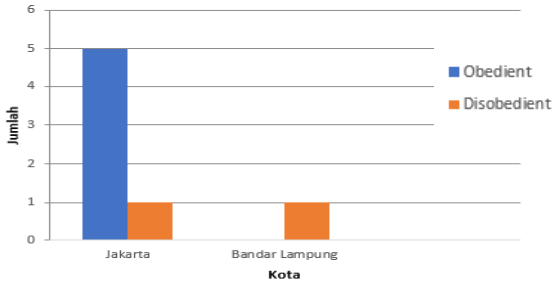


Figure 7. Classification of Overall LSSR Transportation Figure 8. Classification of LSSR Transportation by City

3.1.4 Classification of Physical Distancing Sentiments

Then for the sentiment analysis towards compliance with physical distancing, the results of the being obedient and disobedient were 339 and 80 comments from the overall results which can be seen in Fig. 9. Classification Result of LSSR Physical Distancing Base on City can be seen in Fig.10.

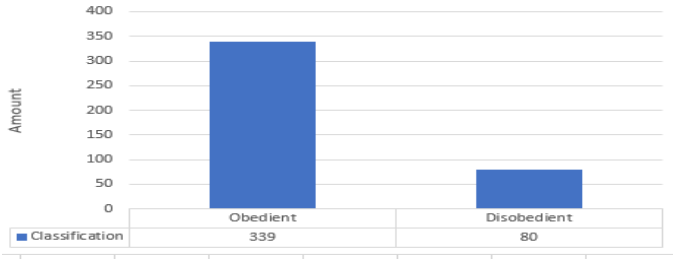


Figure 9. Classification of Overall LSSR Physical Distancing

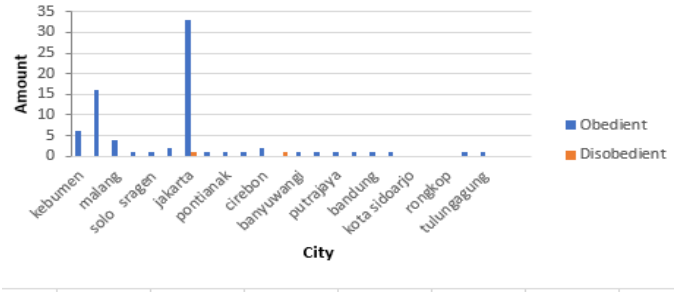


Figure 10. Classification t of LSSR Physical Distancing Base on City

3.2 Interpretation of Knowledge

The interpretations of the resulting knowledge to increase public awareness to better compliance with the regulation regarding LSSR are as follows:

1. Based on Figures 1, 4, 7 and 9 indicate that on average Indonesian people comply with the LSSR rules which consist of LSSR WFH, LSSR study at home, LSSR for transportation and LSSR for physical distancing.
2. From the classification results in Figures 1, 4, 7 and 9, there are still many people do not comply with the LSSR rules, therefore it is necessary to take the following steps to increase community compliance with LSSR:
   a. The government to improve public health awareness and more firm in enforcing regulations related to LSSR.
   b. There must be strict sanctions against people who violate the LSSR rules, to avoid similar mistakes.
   c. The most important thing is that there must be self-awareness from all levels of Indonesian society

3.3 Results Evaluation

The evaluations carried out in this study is Evaluation using the Confusion Matrix. Based on the classification results of each keyword, an accuracy of 98% is obtained for home study, 93% for WFH, 88% for transportation and 90% for physical distancing.

4. Conclusion

The government's efforts to reduce the spread of the Covid-19 outbreak can be carried out through sentiment analysis on the level of public compliance. The main objective of this research is to analyze the sentiment level of community compliance with LSSR activities using the SVM algorithm. In this study, the sentiment analysis of the level of public compliance with the LSSR rules is focused on the rules of studying at home, WFH, transportation, and physical distancing activities. The Indonesian people have the highest compliance with the LSSR physical distancing rules which reaches 80.9%, and the lowest level of compliance with the LSSR transportation rules only reaches 50.21%. The level of compliance of the Indonesian people to the rules of LSSR WFH and studying at home reached 70.7% and 62.67%. This sentiment analysis is also able to map the level of community compliance based on the domicile of cities and provinces. The overall accuracy rate of sentiment analysis reaches 92.25%. This analysis of the level of community compliance can be used to formulate strategies and government policies in the process of reducing the spread of the Covid-19 outbreak, through a process of strengthening socialization and a humanist LSSR control system. A more intensive socialization of LSSR by empowering community strengths who really understand the character of their citizens can be done by referring to a sentiment analysis of the level of community compliance in each city or province.

References:

Apasari P.J. Twitter Sentiment Analysis Using the Lexicon-Based Method and Support Vector Machine. Telkom University. Bandung. (*Analisis Sentimen Twitter Menggunakan Metode Lexicon-Based dan Support Vector Machine*). Universitas Telkom. Bandung, 2017.

Carteni A, Di Francesco L, Martino M. How Mobilit Habits Influenced the Spread of the COVID-19 Pandemic : Result from the Italian Case Study. Science of The Total Environment, Vol. 741, Nov 1, 2020, 140489. https://doi.org/10.1016/j.scitotenv.2020.140489, 2020.

Flynn, M., & Francis, L. *Text Mining Handbrook.* In Casuality Actuarial Society E-Forum, 2010.

Gullo, F. From pattern in data to knowledge discovery: what data mining can do. *Physics Procedia,* 62, pp. 12-18, 2015.

Han, J. & Kamber, M. *Data Mining Concepts and Techniques Second Edition.* Morgan Kauffman, San Francisco, 2012.

Huang, Chia-Hui, Keng-CHieh, Yang, dan Kao, Han-Ying. Analyzing Big Data With The Hybrid Interval Regression Methods. *The Scientific World Journal*, 2014.

Kharde, V., & Sonawane, P. Sentiment Analysis of Twitter Data: A Survey Of Techniques. arXiv preprint arXiv, Vol. 139, 0957-8887, 2016.

Liu, B. Sentiment Analysis and Opinion Mining. Synthesis lectures on human language technologies, 5(1), 1-167. 2012.

Monarizqa, N., Nugroho, L. E., & Hantono, B. S. Application of Sentiment Analysis on Indonesian-Language Twitter as a Rating Officer. (*Penerapan Analisis Sentimen Pada Twitter Berbahasa Indonesia Sebagai Pemberi Rating)*. *Jurnal Penelitian Teknik Elektro dan Teknologi Informasi*, 1, 151–155, 2014.

Nugroho, K. S. Confusion Matrix for Model Evaluation in Supervised Learning.*Confusion Matrix. (Evaluasi Model pada Supervised Learning).* https://medium.com/@ksnugroho/confusion-matrix-untuk-evaluasi-model-pada-unsupervised-machine-learning-bc4b1ae9ae3f, 2019.

Permana, R. H. The 2020 Homecoming Prohibition Applies to LSSR Areas, Here's the List. Retrieved from DetikNews: Larangan Mudik 2020 Berlaku Untuk Daerah LSSR, DetikNews: https://news.detik.com/berita/d-4989468/larangan-mudik-2020-berlaku-untuk-daerah-LSSR-ini-daftarnya/4 (Accessed Mei 16th 2020), 2020.

Saini, S. Sentiment Analysis on Twitter Data using R. *International Conference on Automation*, *Computational and Technology Management (ICACTM)* 68–72, 2019.

Suryahadi, A. The Impact of COVID-19 Outbreak on Poverty: An Estimation for Indonesia. SMERU. Working Paper. The SMERU Research Institute : Jakarta, 2020.

Tosida, E.T., Permana, A., Karlitasari, L., Ardiansyah, D., Andria, F., Bon A.T. Digital Tourism Education Collaboration for Strengthening Micro Business and Post Covid-19 Sustainable Eduacation Models : *Proceedings of the International Conference on Industrial Engineering and Operations Management,* Detroid Michigan, USA, August 10-14, 2020. **pp. 2399-2408, ISSN: 2169-8767.** http://www.ieomsociety.org/detroit2020/papers/489.pdf, 2020a.

Tosida, E.T., Muhaimin, A., Hidayat, M., Ardiansyah, D., Andria, F., Bon A.T. Strengthening the Competitiveness of Micro-Businesses Based on Local Wisdom Through Digital Tourism Education Collaboration *: Proceedings of the International Conference on Industrial Engineering and Operations Management*, Detroid Michigan, USA, August 10-14, 2020. pp. 2439-2446, ISSN : 2169-8767, http://www.ieomsociety.org/detroit2020/papers/494.pdf , 2020b.

Tosida, E. T., Ardiansyah, D., Walujo, A. D. Kujang And Batik Bogor Educational Games To Grow Millennial Generation Enthusiasm For Local Wisdom Through Digital Media. *International Journal of Business, Economics, and Social Development*. https://doi.org/10.46336/ijbesd.v1i2.35, 2020c

Tosida, E. T., Ardiansyah, D., Walujo, A. D., & Sofyandi, A. System Design of Augmented Reality Technology to Strengthen Sustainable Imaging of Kujang Products Based on Local Culture. *International Journal of Recent Technology and Engineering*, *8*(4), 5940–5949. https://doi.org/10.35940/ijrte.d9016.118419, 2019.

Acknowledgements

Acknowledgments are conveyed to the Institute for Research and Community Service of Pakuan University, The Computer Science Department of the Faculty of Mathematics and Natural Sciences, Pakuan University.

Biographies

Eneng Tita Tosida is a lecturer in the Department of Computer Sciences, Faculty of Mathematics and Natural Sciences, Universitas Pakuan. She teaches in Simulation Techniques and Data Mining, Linear Programming and Optimization Models and research methods. She leads research group of Decision Support System (DSS) and Socio Informathic, and actives on educational digital media base on game, Augmented Reality and Virtual Reality research. She also actives on Indonesian Operations Research Association (IORA) as Secretary. Now is serving as head of Community Services Center, Universitas Pakuan.

Erniyati is a lecturer in the Department of Computer Science, Faculty of Mathematics and Natural Sciences, Universitas Pakuan. She teaches in data structure.

Krisna is a graduate Computer Science Department, Faculty of Mathematics and Natural Sciences, Universitas Pakuan.

Abdul Talib Bon is a professor of Production and Operations Management in the Faculty of Technology Management and Business at the Universiti Tun Hussein Onn Malaysia since 1999. He has a PhD in Computer Science, which he obtained from the Universite de La Rochelle, France in the year 2008. His doctoral thesis was on topic Process Quality Improvement on Beltline Moulding Manufacturing. He studied Business Administration in the Universiti Kebangsaan Malaysia for which he was awarded the MBA in the year 1998. He's bachelor degree and diploma in Mechanical Engineering which his obtained from the Universiti Teknologi Malaysia. He received his postgraduate certificate in Mechatronics and Robotics from Carlisle, United Kingdom in 1997. He had published more 150 International Proceedings and International Journals and 8 books. He is a member of MSORSM, IIF, IEOM, IIE, INFORMS, TAM and MIM.