Thai NLP-based Text Classification of the 21st-century Skills toward Educational Curriculum and Project Design

Lawankorn Mookdarsanit[1], Pakpoom Mookdarsanit[2]

[1]Business Information Systems, Faculty of Management Science
Chandrakasem Rajahbat University
Bangkok, Thailand
lawankorn.s@chandra.ac.th

[2]Computer Science and Artificial Intelligence, Faculty of Science
Chandrakasem Rajahbat University
Bangkok, Thailand
pakpoom.m@chandra.ac.th

Abstract— Based on the World economic forum, there are 16 essential 21-century skills that consist of (1) literacy, (2) numeracy, (3) scientific literacy, (4) digital literacy, (5) financial literacy, (6) cultural and civic literacy, (7) critical thinking, (8) creativity, (9) communication, (10) collaboration, (11) curiosity, (12) initiative, (13) persistence, (14) adaptability, (15) leadership and (16) social and cultural awareness. The learning objectives in a curriculum/course or project designed for university students should be fulfilled in some of the 21-century skills. Based on Natural language processing (NLP), this paper proposed a novel Thai text classification for the 21st-century skills that can be seen as Artificial intelligence (AI) in education. An unknown Thai text based on an objective/purpose of the curriculum is classified as the one of those 16 skills using Long-short-term Memory with Attention (ATTN-LSTM). Each Thai word is input to the ATTN-LSTM as learning parameterization. The proposed ATTN-LSTM provides the accuracy higher than 60%. Also, the ATTN-LSTM improves from baseline LSTM as 10% based on our 7,440 raw Thai texts and the attention score helps the correctness classification. ***(Abstract)***

Keywords-Education Technology; Artificial Intelligence in Education; Learning Objective Design; Thai Text Classification; Natural Language Processing;

I. INTRODUCTION

Beyond the educational output, soft (or applied) skills are essential for the student life style [1] in learning, living and working [2]. The skills should be included in all educational curriculums/courses or projects from the university for the student's outcome. As referred to Sustainable development goals (SDGs) [3] with sufficient economy [4], being the development country or developing country depends on human's survival, self-reliance and social compatibility. And the human development totally concerns the education. To that end, education is the main factor for long-term SDGs. There are 16 skills for the 21-century listed by World economic forum [5]: (1) literacy, (2) numeracy, (3) scientific literacy, (4) digital literacy, (5) financial literacy, (6) cultural and civic literacy, (7) critical thinking, (8) creativity, (9) communication, (10) collaboration, (11) curiosity, (12) initiative, (13) persistence, (14) adaptability, (15) leadership and (16) social and cultural awareness, respectively. Since the learning objectives are written in a curriculum or project [6-7] and a textual written objective is often fulfilled in one of those 16 skills, this paper proposes a novel Thai text classification for the 21st-century skills based on Natural language processing (NLP). NLP is proposed to make computer understand the language from text and/or image that can be applied to Artificial intelligence (AI) in education [8-9]. Some Thai NLP applications are machine translation [10], hate speech detection [11-12], COVID-19 fake news detection [13-14], sentiment analysis [15], bully detection [16], automatic image captioning [17], optical character [18]/handwriting [19] recognition and handwriting generation [20-21]. Thai The NLP-based system is based on Long-short-term Memory with Attention (ATTN-LSTM).

From Figure 1, the proposed paper can be abstractly described in 4 following steps.

Step 1: A Thai raw text written in learning objective is input to the system.

Step 2: The system cleans some textual data; and tokenizes all Thai words from the text.

Step 3: Each word is used to compute the Attention score (ATTN) that measures the concerning those 16 skills.

Step 4: Each word with attention score is input to Long-short-term memory (LSTM) to classify the most fulfilled skill from 16 types.

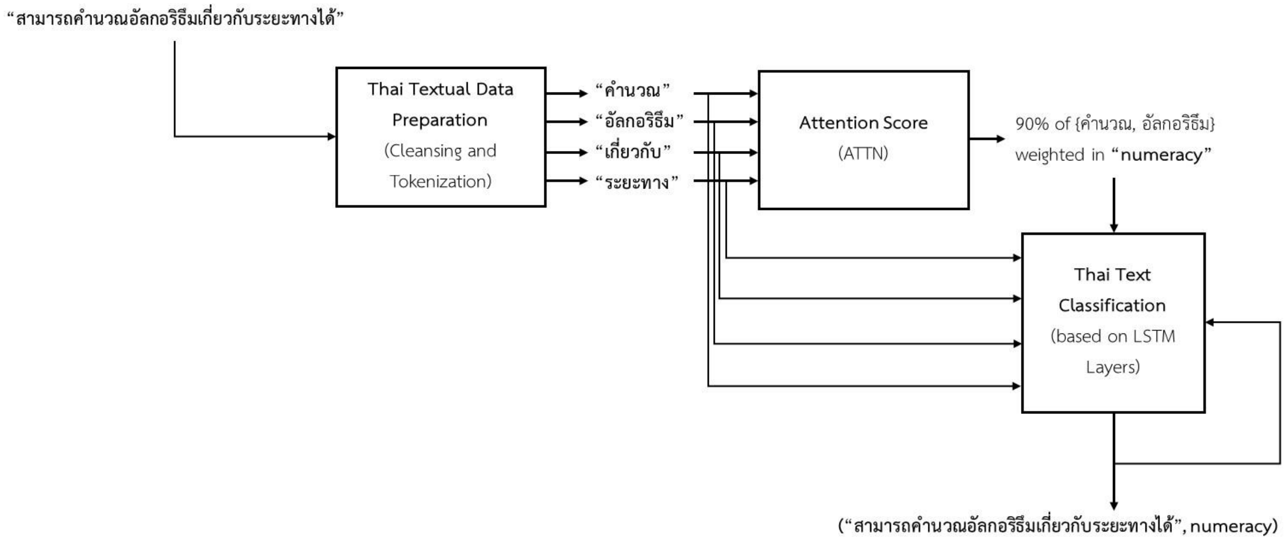
“สามารถคำนวณอัลกอริธึมเกี่ยวกับระยะทางได้”
Thai Textual Data Preparation (Cleansing and Tokenization)
“คำนวณ”
“อัลกอริธึม”
“เกี่ยวกับ”
“ระยะทาง”
Attention Score (ATTN)
90% of {คำนวณ, อัลกอริธึม} weighted in “numeracy”
Thai Text Classification (based on LSTM Layers)
(“สามารถคำนวณอัลกอริธึมเกี่ยวกับระยะทางได้”, numeracy)

Figure 1. The ATTN-LSTM Framework for Thai NLP-based Text Classification of the 21st-century Skills

The contribution of this paper is (1) to apply text classification in 21st-century skills for classifying the learning objective, (2) to introduce a new NLP-based area in learning objective design, (3) to be an Education Technology (EduTech) helping the instructor for curriculum or project design and (4) to improve the baseline LSTM using ATTN-LSTM.

The paper organization can be divided into 5 parts. Thai textual data preparation is in section 2. The section 3 describes attention score. Long-short-term memory (LSTM) and correctness evaluation can be explained in section 4 and 5. And the conclusion is section 6.

II. Thai Textual Data Preparation

This part talks about the Thai raw texts in this work. All Thai texts are randomly from the curriculums/course or projects from 38 Rajabhat universities, Thailand. This part can be divided into 2 sub-parts: (1) Thai textual data collection and cleansing and (2) tokenization.

A. *Thai Textual Data Collection and Cleansing*

All Thai raw texts are learning objectives from the curriculums or projects that have to be fulfilled in one of 16 types from the 21-st century skills. Each text is a single sentence. All textual data collection is crawled during January 2013-December 2020 from Rajabhat universities via open data on the official websites. Rajabhat universities are located in many regions of Thailand that are the important significance to make Thailand become a developed country. There are 7,440 raw texts and tagged the 16 types of 21st-century skills by the expert. Some

B. *Tokenization*

Since Thai is an unsegmented word language, Thai natural language processing (Thai-NLP) in word-level embedding needs tokenization or word segmentation. Tokenization is a local optimization problem in Thai-NLP that is proposed to segment Thai words within a text. Ambiguity can be found in tokenization e.g., “หมากรอบ” can be “หมา-กรอบ” or “หมาก-รอบ”. In this paper, PyThaiNLP API [22] is conveniently used for the tokenization via Python.

III. Attention Score

Attention score (ATTN) [23] is used to define the weight for each Thai word that imply the keyword related to one of the 16 skills for Long-short-term Memory (LSTM). To compute the attention score as Figure 2, the following steps are described.

Step 1: Each word is input as “Input matrix $\begin{bmatrix} x_0 \\ x_1 \\ x_2 \end{bmatrix}$ ”

Step 2: The “Key matrix”, “Query matrix $\begin{bmatrix} q_0 \\ q_1 \\ q_2 \end{bmatrix}$ ” and “Value matrix $\begin{bmatrix} v_0 \\ v_1 \\ v_2 \end{bmatrix}$ ” can be computed by the Matrix input multiplication as (1)-(3)

$$k_i = W_k x_i \tag{1}$$

$$q_j = W_q x_i \tag{2}$$

$$v_i = W_v x_i \tag{3}$$

where x_i as a Thai word input within an Input matrix (an input matrix refers to a Thai text), k_i, q_j and v_i refer to a value from Key matrix, Query matrix and value matrix, W_k, W_q and W_v are the Weight matrices to compute the values

Step 3: The "Attention value" is computed by (4) where D is a dimension and $\sqrt{D}$ is set to 1

$$e_{ij} = \frac{k_i \bullet q_j^{T}}{\sqrt{D}} \tag{4}$$

Step 4: All Attention values are computed by (5) and all of them are summed by (6)

$$a_{ij} = \frac{e_{ij}}{\sum_{i=0}^{n}\sum_{j=0}^{n} e_{ij}} \tag{5}$$

$$y_j = \sum_{i=0}^{n}\left(a_{i,j} \bullet v_i\right) \tag{6}$$

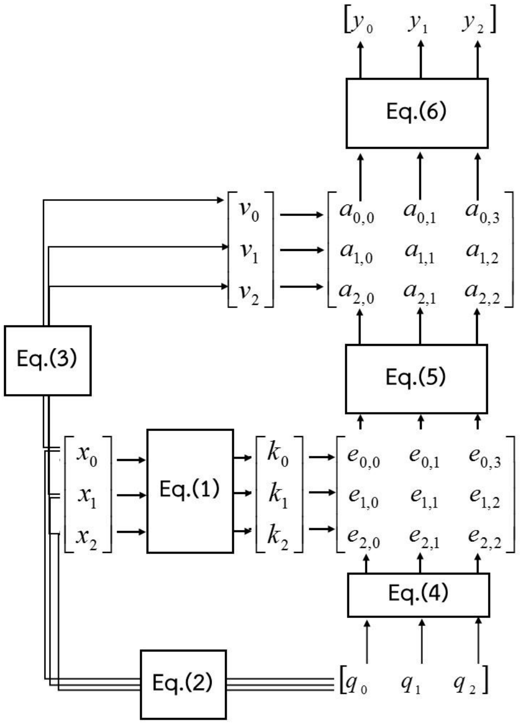
$[y_0 \quad y_1 \quad y_2]$
Eq.(6)
v_0
v_1
v_2
$a_{0,0}$ $a_{0,1}$ $a_{0,3}$
$a_{1,0}$ $a_{1,1}$ $a_{1,2}$
$a_{2,0}$ $a_{2,1}$ $a_{2,2}$
Eq.(3)
Eq.(5)
x_0
x_1
x_2
Eq.(1)
k_0
k_1
k_2
$e_{0,0}$ $e_{0,1}$ $e_{0,3}$
$e_{1,0}$ $e_{1,1}$ $e_{1,2}$
$e_{2,0}$ $e_{2,1}$ $e_{2,2}$
Eq.(4)
Eq.(2)
$[q_0 \quad q_1 \quad q_2]$

Figure 2. Attention computation (ATTN)

IV. Thai Text Classification

Text classification is a research area in Natural language processing (NLP). This paper applies NLP to automatically fulfil the objective learning text in the 16 types based on the 21-century skills defined by World economic forum as a classification problem. This part is divided into (1) supervision and (2) long-short-term memory.

A. Supervision

Supervision or Supervised learning is a machine learning method to build a classifier. Supervision can be divided into 2 portions: training and testing.

(1) Training (or classifier building) is built by textual data and label to teach the computer model. In this case, the textual data is learning objective and the label is the fulfilled 21-st century skills. The combination can be called labeled data.

(2) Testing is a textual data input to classifier and the model automatically adds the label to the data. Adding the label is based on the labeled data during the training.

B. Long-short-term Memory

Long-short-term Memory (LSTM) [24] is a neural network that processes on sequence data. In this paper, the many-to-one LSTM is applied to classify the textual learning objectives that is fulfilled in which the21-st century skills from 16 types.

LSTM is proposed to solve the vanishing gradient in Recurrent neural network (RNN). The architecture of LSTM's hidden layer is shown in Figure 3.

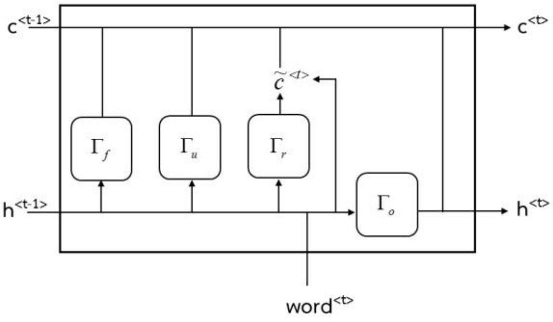

Figure 3. LSTM Layer

Within the LSTM's hidden layer (LSTM Layer) as Figure 3, there are update gate (Γ_u), relevance gate (Γ_r), forget gate (Γ_f) and output gate (Γ_o). The $\tilde{c}^{<t>}$, $c^{<t>}$ and $h^{<t>}$ within LSTM Layer can be shown in (7)-(9).

$$\tilde{c}^{<t>} = \tanh\left(W_c[\Gamma_r \bullet a^{<t-1>}, x^{<t>}] + b_c\right) \tag{7}$$

$$c^{<t>} = \Gamma_u \bullet \tilde{c}^{<t>} + \Gamma_f \bullet c^{<t-1>} \tag{8}$$

$$h^{<t>} = \Gamma_o \bullet c^{<t>} \tag{9}$$

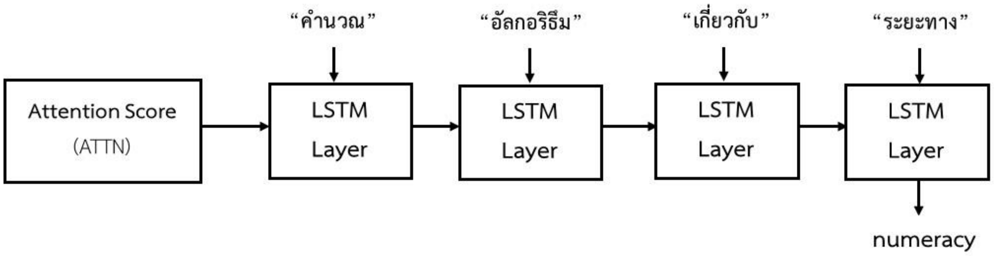

Figure 4. The ATTN-LSTM Framework for Thai NLP-based Text Classification of the 21st-century Skills

V. CORRECTNESS EVALUATION

This part talks about correctness evaluation based on 7,440 raw texts and tagged the 16 types of 21st-century skills that crawled from 38 Rajabhat universities during January 2013-December 2020. The division of this part can be (1) the ATTN-LSTM evaluation and (2) improvement of ATTN-LSTM.

A. The ATTN-LSTM Evaluation

Those Thai learning objective (in form of textual data) with those 16 skills (in form of labels) are trained to the classification model. And some textual data without labels are tested to make the classifier automatically tags the skilling type. Thai text classification in each type is evaluated by recall and precision by (10)-(11).

$$recall = \frac{TP}{TP + FN} \tag{10}$$

$$precision = \frac{TP}{TP + FP} \tag{11}$$

From Figure 5, the highest results in recall are (5) financial literacy and (6) cultural and civic literacy (as 100% recall) from the 21-century skills. The highest precision result is (6) cultural and civic literacy. The 2nd highest precision results are (7) critical thinking and (8) creativity. We have seen that most Text from (6) cultural and civic literacy in curriculums or projects from Rajabhat universities are clear written learning objectives. Moreover, some skills are really similar. For example as (12) initiative, (12) initiative and (8) creativity, (12) initiative and (15) leadership are often mistakes. There are other similarities such as (6) cultural and civic literacy and (10) collaboration. The good objectives written Thai text as examples for the Artificial intelligence will make the model provide the higher correctness.

B. Improvement of ATTN-LSTM

The proposed ATTN-LSTM is also compared to a baseline LSTM. The accuracy can be computed by (12).

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{12}$$

From Table 1, the ATTN-LSTM score improves the accuracy higher than the baseline model as 10%. Since the ATTN helps to compute the weight score before Thai text classification by LSTM, especially in a thousand textual data rows with Thai written-style variability.

TABLE I. TABLE TYPE STYLES

Classifier	**Accuracy**
LSTM	0.57
ATTN-LSTM	0.62

VI. CONCLUSION

This paper applies Thai text classification as one of Natural language processing (NLP) techniques to classify the type of Thai textual learning objective, based on the 21st century skills from Word economic forum. The labeled data is 7,440 raw Thai texts with tagged skills from 38 Rajabhat universities during January 2013-December 2020. The tokenization (or word segmentation) in Thai text is done by PyThaiNLP library. And the baseline Long-short-term Memory (LSTM) can be improved by Attention score (ATTN) as 10%. The correctness of labels in raw texts totally affects the correctness classification. For future work, higher quality of labeling should be provided together with synonyms and antonyms. And the Transformer-based models like BERT, T5 or GPT-3 can synthesize a new written styles. For example, Thai learning objective is input to the model; the model can synthesize a new Thai text styles as an output with the same meaning of textual input.

ACKNOWLEDGMENT

All 7,440 raw Thai texts are from the open data on official Rajabhat universities’ websites. There are 38 Rajabhat universities in many regions of Thailand. According to the Sustainable development goals (SDGs) under the umbrella of United Nation (UN) plan, the power of Rajabhat’s social engineers (Thai: วิศวกรสังคม) can transform Thailand into developed country by the advancement of education technology (EduTech).

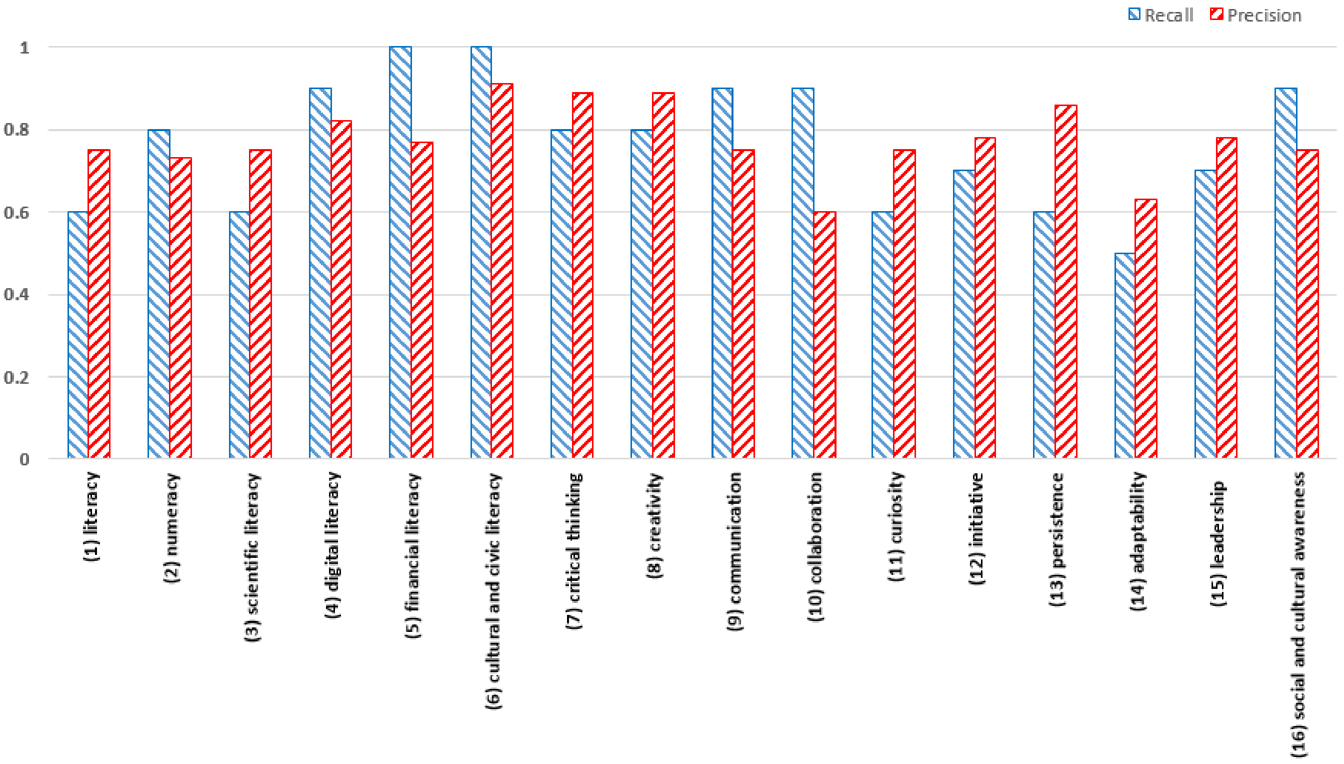

Figure 5. Recall and precision evaluation in Thai text classification based on the 21st-century skills

REFERENCES

[1] M. J. D. Sunarto, "Improving Students Soft Skills using Thinking Process Profile Based on Personality Types," in International Journal of Evaluation and Research in Education, vol. 4, no. 3, pp. 118-129, 2015.

[2] F. F. Patacsil and C. L. S. Tablatin, "Exploring the Importance of Soft and Hard Skills as Perceived by It Internship Students and Industry: A Gap Analysis," in Journal of Technology and Science Education, vol. 7, no. 3, pp. 347-368, 2017.

[3] C. Kroll, A. Warcold and P. Pradhan, "Sustainable Development Goals (SDGs): Are We Successful in Turning Trade-offs into Synergies?," in Nature Palgrave Communications, vol. 5, no. 140, 2019.

[4] P. Barua and P. Tejativaddhana, "Impact of Application of Sufficiency Economy Philosophy on the Well-Being of Thai Population: A Systematic Review and Meta-Analysis of Relevant Studies," in Journal of Population and Social Studies, vol. 27, no. 3, pp.195-219, 2019.

[5] P. J. Puccio, "The Case for Creativity in Higher Education: Preparing Students for Life and Work in the 21st Century," in Kindai Management Review, vol. 8, pp.30-47, 2020.

[6] R. Duarte, A. Lacerda-Nobre, F. Pimentel and M. Jacquinet, "Broader Terms Curriculum Mapping: Using Natural Language Processing and Visual-supported Communication to Create Representative Program Planning Experiences," in arXiv:2102.04811, 2021.

[7] N. Y. Vo, Q. T. Vu, N. H. Vu, T. A. Vu., B. D. Mach and G. Xu, "Domain-specific NLP System to Support Learning Path and Curriculum Design at Tech Universities," in Computers and Education: Artificial Intelligence, vol. 3, pp.1-12, 2022.

[8] L. Chen, P. Chen and Z. Lin "Artificial Intelligence in Education: A Review," in IEEE Access, vol. 8, pp.75264-75278, 2020.

[9] O. Zawacki, V. I. Marin, M. Bond and F. Gouverneur, "Systematic Review of Research on Artificial Intelligence Applications in Higher Education – Where Are the Educators?," in International Journal of Educational Technology in Higher Education, vol. 16, no, 39, 2019.

[10] S. Lyons, "A Review of Thai–English Machine Translation," in Machine Translation, vol. 34, no. 2-3, pp. ,197-230, 2020.

[11] L. Mookdarsanit and P. Mookdarsanit, "Combating the Hate Speech in Thai Textual Memes," in Indonesian Journal of Electrical Engineering and Computer Science, vol. 21, no. 3, pp.1493-1502, 2021.

[12] P. Fortuna and S. Nunes, "A Survey on Automatic Detection of Hate Speech in Text," in ACM Computing Surveys, vol. 51, no. 4, pp.1-30, 2018.

[13] P. Mookdarsanit and L. Mookdarsanit, "The COVID-19 Fake News Detection in Thai Social Texts," in Bulletin of Electrical Engineering and Informatics, vol. 10, no. 2, pp.988-998, 2021.

[14] C. Raj and P. Meel, "ARCNN Framework for Multimodal Infodemic Detection," in Neural Networks, vol. 146, pp.36-38, 2022.

[15] N. N. Alabid and Z. D. Katheeth, "Sentiment Analysis of Twitter Posts Related to the COVID-19 Vaccines," Indonesian Journal of Electrical Engineering and Computer Science, vol. 24, no. 3, pp.1727-1734, 2021.

[16] P. Mookdarsanit and L. Mookdarsanit, "TGF-GRU: A Cyber-bullying Autonomous Detector of Lexical Thai across Social Media," in NKRAFA Journal of Science and Technology, vol. 15, no. 1, pp.50-58, 2019.

[17] P. Mookdarsanit and L. Mookdarsanit, "Thai-IC: Thai Image Captioning based on CNN-RNN Architecture," in International Journal of Applied Computer Technology and Information Systems, vol. 10, no. 1, pp.40-45, 2020.

[18] K. Kesorn and P. Phawapoothayanchai, "Optical Character Recognition (OCR) Enhancement Using an Approximate String Matching Technique," in Engineering and Applied Science Research, vol. 45, no. 4, pp.282-289, 2018.

[19] P. Mookdarsanit and L. Mookdarsanit, "ThaiWrittenNet: Thai Handwritten Script Recognition Using Deep Neural Networks," in Azerbaijan Journal of High Performance Computing, vol. 3, no. 1, pp.75-93, 2020.

[20] L. Kang, P. Riba, Y. Wang, M. Rusinol, A. Fornes and M. Villegas, "GANwriting: Content-Conditioned Generation of Styled Handwritten Word Images," The 2020 European Conference on Computer Vision, Virtual Conference, 2020, pp. 273-289.

[21] L. Mookdarsanit and P. Mookdarsanit, "ThaiWritableGAN: Handwriting Generation under Given Information," in International Journal of Computing and Digital Systems, vol. 10, no. 1, pp.689-699, 2021.

[22] “PyThaiNLP: Tokenization” [Online]. Available: https://pythainlp.github.io/docs/2.0/api/tokenize.html [Accessed: 18 September 2021].

[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser and I. Polosukhin, "Attention is All you Need," The 2017 Conference on Neural Information Processing Systems, Long Beach, California, 2017.

[24] S. Hochreiter and J. Schmidhuber, "Long Short-term Memory," in Neural Computation, vol. 9, no. 8, pp.1735-1780, 1997.