Sentiment analysis of messages on Twitter related to COVID-19 using deep learning approach

Chotika Imvimol
Computer Engineering Department,
Faculty of Engineering
Chulalongkorn University
Bangkok, Thailand
6272029521@student.chula.ac.th

Prabhas Chongstitvatana
Computer Engineering Department,
Faculty of Engineering
Chulalongkorn University
Bangkok, Thailand
prabhas.c@chula.ac.th

***Abstract*—– The widespread situation of the Coronavirus-19 (COVID-19) pandemic is a tangible and pressing concern. Many changes in terms lifestyle are necessary to reduce the chance of infection. While citizens have gone through different emotions, they would share their thought and interactions on social media, especially on Twitter. COVID-19 related messages can imply social emotion. This study performs sentiment analysis on tweets and annotated them into six classes of positive and negative feelings consist of anger, disgust, fear, sadness, joy, and surprise. We analyzed both textual information and historical data. We collected 120,642 unique tweets datasets between 1 January 2020 and 30 June 2021. We compared the performance of five neural network models which are multi-layer perceptron, RNN, LSTM, Bidirectional LSTM, and GRU with several metrics consists of accuracy, F1 score, precision, and recall. The results show that LSTM model has the highest accuracy score at 77.4% while GRU has the best F1 score at 77.13%. These models could be used to monitor the movement of negative emotions. In addition, we provide interesting insights from sentiment analysis with tweet data and historical reported of infected cases, and vaccination data.**

Keywords— Sentiment, Classification, Deep Learning, COVID-19, Twitter

I. INTRODUCTION

Coronavirus Infection Disease 2019 (COVID-19) has been spreading the world. This virus can easily infect people and it has effects from mild symptoms to severe. People need to be aware of it and live with it until the situation improved. The situation frequently affects people in bad ways especially emotion. People use social media to express, and share thought with other people including the COVID-19 topic. This could imply that emotions are related with the situation.

In this situation, people express their feeling in three types which are positive, neutral, and negative. Moreover, every type can divide into different levels of emotion, it can separate into six simple categories consist of anger, disgust, fear, sadness, joy, and surprise. The more specific emotion messages are divided, the more information can be studied. These five groups of emotions would be provided deeper insights compared with three categories of data. Studying those messages can help to understand public emotions during a specific situation. Classify COVID-19 related messages on social media would use to monitor emotion and provide new information for policy makers to solve public emotion problems. In addition, analyzing data in Thailand focusing would bring the more useful relationship between infection situations with people via their expression.

A. Coronavirus-19

COVID-19 first appeared in December 2019 and World Health Organization (WHO) reported it on 11 March 2020 in Wuhan, China and after that, the number of patients cases was burst up in a short period in the local area. A month after, it spreads out aboard and the situation became worse in the rising of new cases. People in many countries around the world suffered from it especially the USA, India, Brazil, and UK because they have a much higher number of the patients.

In Thailand, the first found COVID-19 cases on 12 January 2020, after that the situation continually is getting worse because of the limit of the capacity of healthcare services to cover active cases. There are significant increases in the number of Thai patients since the end of the first quarter of 2021 and became a crisis in August 2021 when the month of the highest rise of new cases are reported. After that, the situation in Thailand is getting better than in the mid of the year.

There are several symptoms from this virus, for instance, fever, cough, loss of taste or smell, sore throat, aches and pains, diarrhea, difficulty breathing, loss of mobility or speech even dead. People need to be aware of these symptoms and take precautions to reduce the possibility of infection.

B. Twitter

Twitter is a social media platform where people can connect and communicate with short messages called tweets. The posting can be of various types which are 280 characters or less text, link, picture, and a short video. Users can react with the posting by comment, like, retweet, and sharing to friends or another platform. Users can add hashtag character (#) into the context in the post to follow the topic easier and if many people post a lot about the topic in the same hashtag, it will show up on the top chart to inform that people are interested in this topic at that time.

Many Thai Twitter users share and express their opinions on many topics, including COVID-19 in Thailand. They share about infection situations, their emotion-related to COVID-19, and other stories.

C. Deep learning

Deep learning is a class of algorithms that have multiple layers of neuron structure like a human brain. The algorithms find the pattern in data by learning from examples. The hierarchy of algorithm learn input data and provide output. It can achieve many tasks, for example, prediction, classification, and time series analysis. It can handle both supervised and unsupervised learning task.

Our study collects tweet messages related to COVID-19 to classify emotion on Thai language dataset from Twitter users. There are six classes of positive and negative feelings. We use WangchanBERTa to label data since it was trained on the same domain of data with our data. After that, the classification task is done with the neural network technique. Finally, providing data analysis and visualization messages and infection situation trends and news.

The rest of this paper is organized as follows. Firstly, Section II is related work. Section III gives information about the dataset. Section IV, V, and Section VI describe the method, experiment results, and discussion. Finally, Section VII gives the conclusion.

II. RELATED WORK

There are several researches on emotion classification on COVID-19 related tweet messages. Al-Laith A. and Alenezi M. [1] studied sentiment and symptom classification from January to August 2020 on Arabic language dataset. They labeled some datasets by manual labeling and train Fasttext model which is a neural network to label the rest. Their experiment on LSTM model achieves 82.9% on F1 score. Moreover, they provide topic detection to find the important word in the dataset. Garcia K. and Berton L. [2] classified sentiment and topic detection on both English and Portuguese language in USA and Brazil between April and August 2020 by using an auto-labeling library to create labels. They found that the best approach is logistic regression and Linear SVM which provide performance equal to 87% on F1 score. Kausar M.A., et. al., [3] reported emotional classification in many countries around Persian Gulf area between 21 June and 20 July 2020 with 50000 tweets using automate library to prepare a dataset and classify into 3 classes which are positive, neutral, and negative feeling. Mathur, A. et. al., [4] studied on an English language dataset from 22 January to 15 April 2020 contains 30,000 tweets and classified into eight classes which are anger, hope, disgust, fear, positive, negative, joy, sadness, and surprise. Pasupa, K. and Ayutthaya, T.S.N. [5] studied sentiment analysis on Thai children tales dataset using different in word embedding for semantic, POS-tag for grammar and emotion of word into emotional score. They compared result on deep learning models which are CNN, LSTM, BLSTM. The result showed CNN achieve highest F1-score at 81.7% Pasupa, K. and Ayutthaya, T.S.N. [6] trained many models which were fundamental deep learning consists of CNN and LSTM and hybrid models as BLSTM-CNN, CNN-BLSTM, BLSTM+CNN, and BLSTM×CNN on Thai-SenticNet5 corpus into positive, neutral and negative feelings and evaluated on three Thai language social media datasets which were ThaiTales, ThaiEconTwitter, and Wisesight and The highest performance was BLSTM-CNN on F1-scores at 74.36%, 77.07%, and 55.21%, respectively.

III. DATASET

Our data is collected from Twitter API called Twint library via many hashtags consist of 9 hashtags which are COVID, COVID19, COVIDtoday, COVID19today, VirusCorona2019, VirusCorona, CoronaViras, Corona. Fig. 1 shows translation of hashtags. The data of this experiment are Tweets related to COVID-19 in Thai language between 1 January 2020 and 30 June 2021. The unique message in the dataset consists of 120,642 tweets. Moreover, the official reported data of new infection patients and vaccination people in Thailand via DDC open data API which is in the authorization of the Department of disease control, Ministry of Health in Thailand.

โควิด	COVID
โควิด19	COVID19
โควิดวันนี้	COVIDtoday
โควิด19วันนี้	COVID19today
ไวรัสโคโรนา2019	VirusCorona2019
ไวรัสโคโรน่า	VirusCorona
โคโรน่าไวรัส	CoronaViras
โคโรน่า	Corona

Fig. 1. Translation of hashtags from Thai to English

IV. MEthod

Our experimental start from preparing, annotating, selecting candidate deep learning models to classify datasets base on the fundamental based model which is multi-layer perceptron, and recurrent based models which are RNN, LSTM, and GRU to get the best fit model. Lastly, finding insights from input with Thailand infection situation.

A. Data Preprocessing

The first step is to clean data by removing punctuations, tabs, blank space, numbers, hashtags, user mentions, emoticons, and special characters such as "#" and "@" in the messages. The next process is to obtain, unique messages for classification and data analysis purpose

B. Data Labelling

Since the size of dataset is too large to do hand labeling, we use WangchanBERTa pre-train model to predict 6 classes of label for our dataset. There is around 10-15% non-accurate label compares with native judgment. We validate the transformer mechanism-based label by hand using a random pick sample to prove around 3,000 tweets.

C. Word Tokenize

This work used Attacut approach which is a fast and accurate neural network-based Thai word segmenter to cut sentences into a single word developed by Chormai, P. et. al., [11].

D. Feedforward neural network

The simplest structure of the neural network is Feedforward Neural Network. The information is always fed in the forward direction. The network does not have any loops or cycles. The network consists of the input layer, many interconnected in the feed-forward hidden layer which is called multi-layer perceptron and output layer. Each layer has input, hidden and output nodes with an activation function.

Multi-layer perceptron networks mostly use back-propagation method to learn data by applying the nonlinear transformation function to every neuron, create output, compare the output with the correct answer to calculate predefined error-function. The error that the network gets is fed back to adjust weights to decrease error while learning. The algorithms continue to learn from datasets until they can get satisfying results.

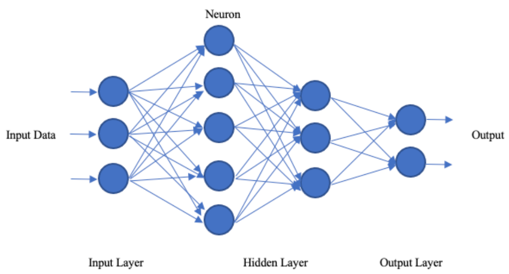

Fig. 2. Multi-layer perceptron networks

E. Recurrent neural network (RNNs)

RNNs are a subclass of neural networks which can handle sequential data or time-series data since they can remember prior inputs and use them to influence current input and output. This type of architecture performs well in speech recognition and natural language processing task.

In Fig. 3, Rolled RNN represents one-time step of this network while unrolled RNN picture show many times and use hidden layers to learn data as a sequence. Each layer of RNN uses the same weight for entire of the network. Weights are adjusted through the learning process by backpropagation method. However, the learning process of RNNs suffered from vanishing gradients when the size of the gradient is too small which means the model stops learning.

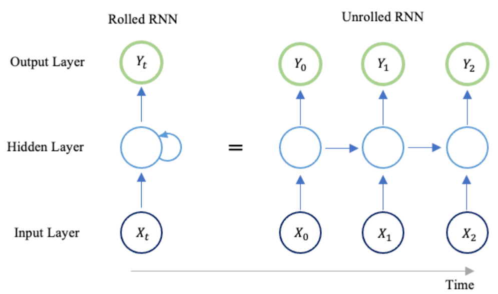

Fig. 3. Recurrent neural network

To overcome this problem, the modified version of RNN, long short-term memory (LSTMs) was introduced by Hochreiter, S. et. al., [12] They create new cells in hidden layers, and there are three gates which are input, forget and output gate that use a suitable amount of previous data to predict output in the next step. In addition, combining with bidirectional to feed data into networks, resulting in the bidirectional model. The prediction of this model works well for a sequence of text input, but they cannot be used for time series prediction because prediction problems cannot let the model know future data.

Gated recurrent unit (GRUs) is the smaller version of LSTMs architecture because they have redesigned three gates to two gates which are reset and update gates to overcome a series of data problems. The gates use to control information issues. The size of this model is smaller and lighter than LSTMs. (fig. 4)

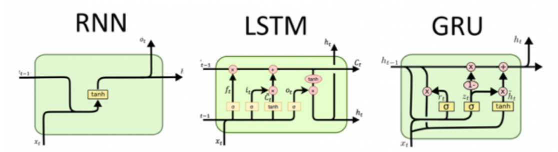

Fig. 4. The cell in RNNs, LSTMs, and GRUs
Source: Figure from [13]

F. Transformer

A transformer is an encoder-decoder-based neural network with an attention mechanism that determines the important relationship of input, for instance, collocation of words that both words are likely be used to together. The transformer was proposed in 2017 by Brain A. V. G. et. al., [14]. This mechanism can solve sequences problem well, especially on text data, leading to the popularity for solving natural language processing problems. (Fig. 5)

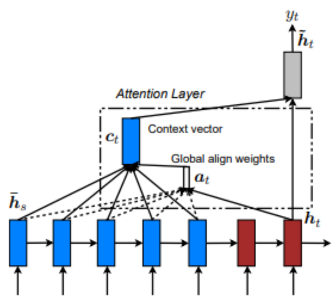

Fig. 5. Attention layer
Source: Figure from [15]

Bidirectional Encoder Representations from Transformers (BERT) is a language model which represented the relationship of the word in language in terms of statistic numbers. BERT model consists of the encoder from the transformer-based technique with attention, feed-forward layers. It was proposed by Devlin J. et. al., [16]. BERTs are the state-of-the-art in natural language processing in these decades because they can achieve most tasks in greater performance than other models or techniques.

WangchanBERTa is pretraining transformer-based Thai Language Models. It was proposed by Lowphansirikul L. et. al., [17]. They provided a model of Thai language which was trained on a large Thai language dataset that has a total size of 78 GB in several domains such as social media messages, news articles, and other public datasets. Their models outperform strong baselines in many downstream tasks such as sequence classification and token classification.

We load the pre-train dataset from WangchanBERTa for annotating data using Transformers version 3.5.0, Thai2transformers version 0.1.2, and Pytorch version 1.4.0. and training neural network model. We use an open-source system library called Keras for deep learning framework with TensorFlow version 2.6.0 as the backend and Python version 3.7.11. We divide the dataset into train, validation, and test dataset. There are 96,513 messages (80% of all datasets) which are divided into 77,210 messages (80%) for train dataset and 19,303 messages (20%) for validation dataset, and the rest is test dataset 24,129 messages (20%).

Neural network models in this study are multi-layer perceptron, RNN, LSTM, bidirectional LSTM, and GRU with sequence, embedding, dense, fatten layer. We fine tuning models until they converged to get highest performance by grid search technique. The greatest parameters for our models are Adam optimizer and categorical cross-entropy for multilabel classification with a learning rate 0.001. The model is fitted with 64 batch sizes and train each model on 15 epochs. Fig. 6 shows configuration of LSTM model.

G. Model Comparison

We perform experiments on five neural network models which are multi-layer perceptron, RNN, LSTM, Bidirectional LSTM, and GRU to find the best model for emotional classification.

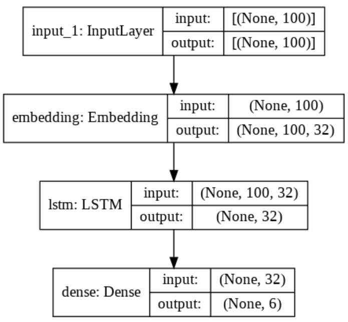

Fig. 6. LSTM Model Architecture

V. Results

The performance on five neural network models is reported is Table I. The performance metrics are accuracy, F1 score, precision, and recall.

TABLE I. Model performance

Model	*Accuracy*	*F1 Score*	*Precision*	*Recall*
Multi-layer perceptron	44.67	42.77	47.09	39.29
RNN	75.24	74.88	76.50	73.39
LSTM	77.40	76.97	78.78	75.29
Bidirectional LSTM	76.33	75.94	77.72	74.30
GRU	77.28	77.13	78.57	75.79

The best performance in our study is LSTM and GRU model which outperforms other models on these metrics. LSTM provides the highest accuracy at 77.40% and precision score at 78.78% while GRU has the best F1 score at 77.13% and recall at 75.79%. RNN based models perform better than the simple neural network. The recurrent based fit well on this data since the complex of structure mechanism and the memory of model are fit well with contextual data and sequences of word or sentences. The performance of LSTM is not significantly higher than BLSTM. It may come from the short text input and the complex structure of BLSTM, resulting in closely in their performance result. There did not have the greatest significant model to beat among others then we can pick the best models via specific metrics for example when we want model that predict sample rights more than try to predict cover all of sample then we should use LSTM, On the other hand, we can use GRU not only the smaller size of model but also people want to get high performance in recall metric which measure how much model can predict cover the right samples.

VI. Discussion

The significant relationship between sentiment messages with COVID-19 infection situation in Thailand can describe in data analysis and visualization as follow.

A. Emotion movement during COVID-19

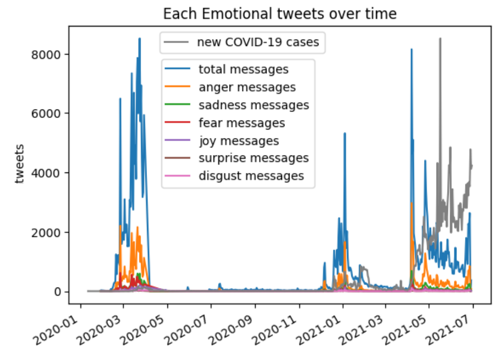

Fig. 7. Emotion movement during COVID-19 pandemic

Emotion during the COVID-19 pandemic of Thai language Twitter users was at the peak 3 times in a period. The first peak in the number of emotional messages around February to April 2020. It was during this first period of this virus spreads that there is news about running out of face masks in Thailand [18], [19]. Following between December 2020 and January 2021, users feel angry with the finding of huge cluster in Shrimp Samutsakhon Market and Lumpini Boxing Stadium [20],[21]. Lastly, the third peak of emotional messages from users around April to May 2021 was during the sharp rise in the number of new cases and death mainly in Bangkok and from prison [22],[23].

B. Anger emotion with official records of infection cases

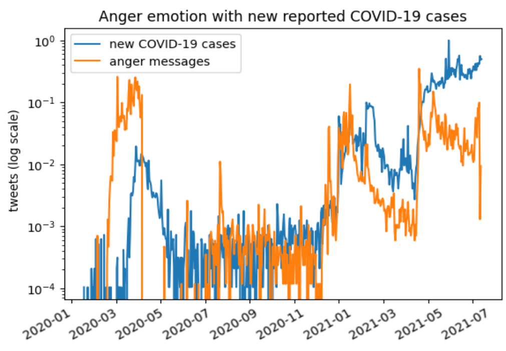

Fig. 8. Anger emotion related with COVID-19 new patients

Plotting Anger emotion which have significant relationship with real situation of new reported cases during COVID-19 infection trend compare with daily reported cases using normalizing data into same scale and plot in log scale. The figure shows most of the anger feelings are fluctuating in the same period except the first and last quarter of entire of the period.

C. *Fear emotion with offical records of vaccinated people*

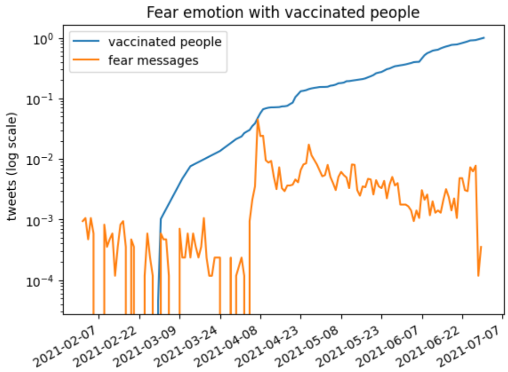

Fig. 9. Fear emotion related with vaccinated people

Plotting fear emotion which have strong relationship with vaccination trend by normalizing data into same scale and plot in log scale. The figure shows most fear feelings are fluctuating up and down between February to April 2021. After that, the number jumped up in a short period and it was likely to be a downward trend in the late of April during the rise in the progress of vaccination people.

VII. Conclusion

In this paper, sentiment analysis of Thai language messages from Twitter COVID-19 related was reported. The results show that the recurrent neural network-based performs well on COVID-19 related messages, especially LSTM which has the best accuracy at 77.40% while GRU has the highest F1 score at 77.13%. Moreover, we provide interesting insight via data analysis from trends comparing with the historical number of daily reported cases and vaccinated people. Further work is finding a better way to annotate data, implement another mechanism to solve this task, and improving the performance of the model.

References

[1] Al-Laith, A. and Alenezi, M., "Monitoring people's emotions and symptoms from Arabic tweets during the covid-19 pandemic," *Information*, vol. 12, no. 2, p. 86, 2021.

[2] Garcia, K. and Berton, L., "Topic detection and sentiment analysis in Twitter content related to covid-19 from Brazil and the USA," *Applied Soft Computing*, vol. 101, p. 107057, 2021.

[3] Kausar, M. A. et. al., "Public sentiment analysis on Twitter data during COVID-19 outbreak," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 2, 2021.

[4] Mathur, A. et. al., "Emotional analysis using Twitter data during pandemic situation: Covid-19," *2020 5th International Conference on Communication and Electronics Systems (ICCES)*, 2020.

[5] Pasupa, K. and Ayutthaya, T.S.N., "Thai sentiment analysis with deep learning techniques: A comparative study based on word embedding, POS-tag, and sentic features". *Sustainable Cities and Society*, *50*, p.101615, 2019.

[6] Pasupa, K. and Ayutthaya, T.S.N., "Hybrid deep learning models for thai sentiment analysis". *Cognitive Computation*, pp.1-27., 2021.

[7] Mackey, T. et. al., "Machine learning to detect self-reporting of symptoms, testing access, and recovery associated with covid-19 on Twitter: Retrospective big data infoveillance study," *JMIR Public Health and Surveillance*, vol. 6, no. 2, 2020.

[8] R. Singh and R. Singh, "Applications of sentiment analysis and machine learning techniques in disease outbreak prediction – A Review," *Materials Today: Proceedings*, 2021.

[9] Müller, M. et. al., "Covid-Twitter-Bert: A natural language processing model to analyse COVID-19 content on Twitter," *arXiv.org*, 15-May-2020. [Online]. Available: https://arxiv.org/abs/2005.07503. [Accessed: 12-Oct-2021].

[10] Naseem, U. et. al., "Covidsenti: A large-scale benchmark Twitter data set for covid-19 sentiment analysis," *IEEE Transactions on Computational Social Systems*, vol. 8, no. 4, pp. 1003–1015, 2021.

[11] Chormai, P. et. al., "Attacut: A fast and accurate Neural Thai word segmenter," *arXiv.org*, 16-Nov-2019. [Online]. Available: https://arxiv.org/abs/1911.07056. [Accessed: 12-Oct-2021].

[12] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[13] dProgrammer lopez, "RNN, LSTM & GRU", 6-April-2019. [Online]. Available: http://dprogrammer.org/rnn-lstm-gru. [Accessed: 12-Oct-2021].

[14] Brain, A. V. G. et. al., "Attention is all you need," *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 01-Dec-2017. [Online]. Available: https://dl.acm.org/doi/10.5555/3295222.3295349. [Accessed: 12-Sep-2021].

[15] Sharma, A. et. al., "Attention-based Neural Machine Translation", 10-March-2019. [Online]. Available: https://towardsdatascience.com/attention-based-neural-machine-translation-b5d129742e2c. [Accessed: 12-Oct-2021].

[16] Devlin, J. et. al., "Bert: Pre-training of deep bidirectional Transformers for language understanding," *arXiv.org*, 24-May-2019. [Online]. Available: https://arxiv.org/abs/1810.04805. [Accessed: 12-Sep-2021].

[17] Lowphansirikul, L. et. al., "Wangchanberta: Pretraining transformer-based Thai language models," *arXiv.org*, 20-Mar-2021. [Online]. Available: https://arxiv.org/abs/2101.09635. [Accessed: 12-Sep-2021].

[18] The standard, "January 12, 2020 – found the first COVID-19 patient in Thailand.", 12-Jan-2021. [Online]. Available: https://thestandard.co/onthisday120163. [Accessed: 12-Oct-2021].

[19] BBC Thai News, "Coronavirus: Situation of medical masks in Thailand, Ran out of it-Free- Complimentary-Recycle", 3-Mar-2021. [Online]. Available: https://www.bbc.com/thai/thailand-51717574. [Accessed: 12-Oct-2021].

[20] BBC Thai News, "Covid-19: Closed the shrimp market in Mahachai, Samut Sakhon province, after three days, found infected patients from the local area in Thailand.", 18-Dec-2020. [Online]. Available: https://www.bbc.com/thai/thailand-55359016. [Accessed: 12-Oct-2021].

[21] Thairath News, "Throwback to the time found a large number of people who infected from the boxing area in Thailand.", 30-Dec-2020. [Online]. Available: https://www.thairath.co.th/sport/others/2003033. [Accessed: 12-Oct-2021].

[22] The Bangkok insight, "Covid-19 situation update on 4 May 2021", 30-Dec-2020. [Online]. Available: https://www.thebangkokinsight.com/news/politics-general/politics/612936/. [Accessed: 12-Oct-2021].

[23] BBC Thai News, "COVID-19: What is causing the coronavirus outbreak in prisons? After the number of infected people surpassed 10,000.", 17-May-2020. [Online]. Available: https://www.bbc.com/thai/thailand-57144771. [Accessed: 12-Oct-2021].