P-ISSN 2586-9000
E-ISSN 2586-9027
Homepage : https://tci-thaijo.org/index.php/SciTechAsia
Science & Technology Asia
Vol.29 No.2 April - June 2024
Page: [138-147]
Original research article

Developing Multi-Label Classification Model for Improving Text Categorizing Problems a Case of Traffy* Fondue Platform

Phopthorn Kaewvichit[1], Akkaranan Pongsathornwiwat[2,*]

[1]*Data Analytics and Data Science, Graduate School of Applied Statistics, National Institute of Development Administration (NIDA), Bangkok 10240, Thailand*
[2]*Logistics Management, Graduate School of Applied Statistics, National Institute of Development Administration (NIDA), Bangkok 10240, Thailand*

Received 15 January 2024; Received in revised form 29 February 2024
Accepted 1 April 2024; Available online 25 June 2024

ABSTRACT

This paper develops an integrated deep-learning-based model for a multi-label text classification problem in order to enhance the efficiency of Traffy* Fondue which is a comprehensive platform for diverse issues of citizen's complaints over the Bangkok metropolitan area. The dataset of Traffy* Fondue has been found to have inaccuracy problems of label categorization, especially in the 'Others' category, which results in delaying of coordination and problem-solving to address the complaint on time. To overcome this problem, this study develops an integrated deep-learning-based model for multi-label text classification problem. Five main methods have been applied to model text classification, namely, Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), Bi-LSTM, fastText and WangchanBERTa. The performance of developed models has been tested against traditional algorithms. The result of modeling in the 'Others' dataset shows that the Bi-LSTM-, CNN+Bi-LSTM- and WangchanBERTa-based approaches are the best three models that outperform others with higher precision, recall, and F1 score. Our approach offers a promising solution to expedite issue resolution and improve coordination within Bangkok's civic management framework.

Keywords: Deep learning; Multi-label classification; Natural language processing; Public complaint data; Text categorization

***Corresponding author:** akkaranan@as.nida.ac.th
doi: 10.14456/scitechasia.2024.32

1. Introduction

The Traffy* Fondue platform is a comprehensive platform which is designed to efficiently manage and address various issues that arise from those who live in the Bangkok metropolitan area. From reporting repairs in office buildings to filing complaints, incidents, disasters, and even clues about corruption, this platform serves as a valuable tool for service departments to tackle problems promptly and effectively. By providing detailed information, including images and locational coordination, staff members are equipped to make informed decisions and expedite solutions.

In operations, the platform categorizes complaint problems into 25 distinct categories. The proposed categories are suited well in real situation; however, users may encounter challenges in selecting the most suitable category. Once facing this situation, they usually fill the complaint out to the “Other” category. This issue leads to delays in problem resolution by the officers. In addition, the "Other" category represents a significant issue of inconsistently ranking which is shown in a problem report across the platform. Addressing problems within this category requires special attention, contributing to additional delays in coordinating efforts among various assigned units.

As observed, the platform's challenge lies in the diverse nature of the categorized issues, creating difficulties for users in accurately selecting the appropriate category. This complexity results in a bottleneck effect, hindering the swift resolution of problems. Additionally, the prevalence of issues falling under the "Other" category underscores the necessity for meticulous effort in categorizing these problems correctly.

In response to these challenges, this research presents an innovative deep learning model for multi-label text classification. The primary objective is to accurately categorize the reported problems, including those labeled as "Others," with the ultimate goal of enhancing platform efficiency and reducing coordination delays when addressing various problems within Bangkok

2. Theory and Related Works

2.1 Text classification

Text classification is a popular task in machine learning. In the traditional approach, examples are associated with a single label from a set of distinct labels. Binary classification involves two labels [1], while multi-class classification deals with more than two. For instance, a recent study tackled the challenge of efficiently handling vast internet information through Chinese text classification. They proposed a comprehensive method combining Word Autoencoder (WAE) and Support Vector Machine (SVM) algorithms. By introducing the WAE topic model, they observed enhanced accuracy and reduced processing time in their experiments [8].

Similarly, in another investigation focused on Turkish texts, the classification performance of various algorithms was compared. Results showed that Multinomial Naïve Bayes, Bernoulli Naïve Bayes, Support Vector Machine, K-Nearest Neighbor, and Decision Trees achieved around 90% classification success [9]. Another researcher, Bo Zhang, employed deep learning models to automatically extract features for news text classification, leading to a 20% improvement in effectiveness compared to SVM when handling short texts [10].

In a different context, Yao T et al. proposed a text classification model based on Fast-Text to overcome limitations faced by traditional machine learning algorithms. The model's application in classifying the emotional polarity of user comments on the Baidu Dianshi platform yielded an accuracy of 0.9275 [11].

Overall, these studies showcase the significance of text classification in various languages and demonstrate the efficacy of different methodologies to achieve accurate and efficient classification results.

2.2 Multi label classification

In the realm of text analysis, multi-label classification emerges as a pivotal task, aimed at effectively assigning textual content into distinct and relevant categories. This task becomes particularly pertinent when each message holds the potential to align with one or more labels, allowing for a nuanced classification approach.

Table 1. The example of multi – label classification problem.

Binary		Multi-class		Multi-label	
X	label	X	label	X	label
x_1	0	x_1	0	x_1	[0,1,2]
x_2	1	x_2	1	x_2	[1,2,3]
x_3	1	x_3	2	x_3	[1,3,4]
x_4	0	x_4	1	x_4	[0,1,3]
x_5	1	x_5	2	x_5	[1,2,3]

For example, this study is centered on the categorization of extensive student remarks obtained from a university in China, specifically focusing on the theme of "engineering sustainability." These remarks delve into subjects such as self-perception, life perspectives, and the comprehension of sustainable development. The comment lengths vary between 500 and 900 characters. In pursuit of this investigation, two distinct deep learning algorithms were employed: BERT and a novel method called MLformer.

The analysis reveals that the MLformer approach outperforms BERT, achieving an accuracy rate of 0.80, while BERT attains an accuracy of 0.75 [7]. Additionally, in the realm of deep learning for multi-label text categorization, Zhang Q et al. applied a TextCNN-based strategy to classify fault text data emanating from power terminal devices [3]. Moreover, Alsukhni B tackled the challenge of multi-label classification for Arabic news through the utilization of MLP and LSTM [4].

In a similar vein, Kim D et al. introduced EnvBERT model (KoBERT) pre-trained with Korean text data, a multi-label text classification model constructed upon the foundations of BERT. Notably, EnvBERT effectively handles issues inherent in imbalanced and noisy environmental news data. By integrating multi-label characteristics and implementing oversampling and threshold techniques, EnvBERT demonstrates a noteworthy enhancement of over 80% in classification accuracy for such demanding datasets [6].

To further enhance the model's efficacy, several research endeavors have proposed innovative algorithms to amplify its performance. Wang S et al. addresses the limitations of conventional CNN and RNN models in preserving crucial text features during feature extraction. They introduce a hybrid model that amalgamates multi-channel CNN with LSTM to proficiently capture both local and long-range features. This incorporation of multi-channel CNN with BERT substantiates improved classification performance, yielding augmented accuracy and affirming its potency in feature extraction [5].

In another notable advancement, Jia L et al. present a novel approach to multi-label classification, leveraging joint neural networks and label correlations (MLNNC). This methodology encompasses nonlinear feature mapping for input data, the acquisition of label-specific features, and the integration of label correlations to refine the model and enhance classification accuracy. Empirical outcomes underscore the efficacy of this approach, as it surpasses prevailing algorithms, although further fine-tuning is warranted for multi-label datasets with limited samples [2].

2.3 Deep learning for text classification

2.3.1 Convolutional Neural Network (CNN) text classification model

CNNs, known for their success in image processing, have also proven effective in text classification. The model employs convolutional layers to extract hierarchical features from input text. CNNs can capture local patterns and dependencies within sentences, making them suitable for tasks such as identifying key phrases and document categorization.

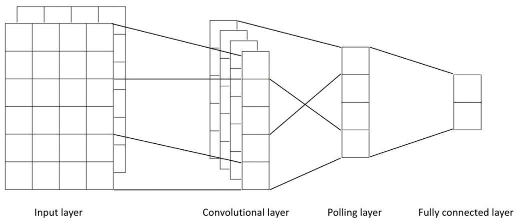

Fig. 1. The structure of TextCNN model.

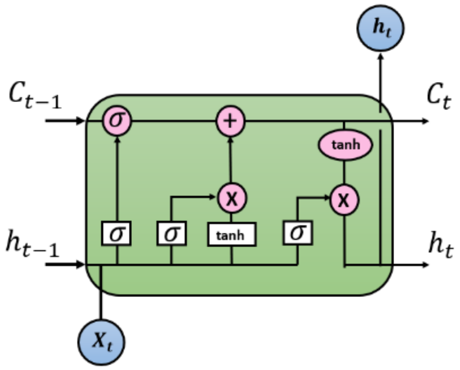

Fig. 2. The structure of Long short-term memory model.

2.3.2 Long Short-Term Memory (LSTM) text classification model

LSTMs, a type of recurrent neural network (RNN), are designed to capture long-term dependencies in sequential data. In the academic domain, LSTMs excel at understanding context and relationships within lengthy paragraphs or documents. This makes them well-suited for tasks requiring the analysis of nuanced language and complex writing structures.

An LSTM employs a specialized memory unit to manage the retention and removal of information [14]. Typically comprised of four integral components, the LSTM structure is designed for effective information flow control [15]. The details of each integral components are explained in the following.

Input Gate: This gate governs the influx of information into the cell state, determining which details from the input should be stored in the cell state. Activation of the input gate is orchestrated by a sigmoid function.

Forget Gate: Responsible for deciding which information in the cell state should be retained or discarded, the forget gate utilizes the prior cell state and the current input to produce a forget factor for each element in the cell state. The forget gate is also regulated by a sigmoid activation function.

Cell State (Memory Cell): Operating as a long-term memory unit capable of storing information across extended sequences, the cell state runs vertically along the entire LSTM chain. It possesses the ability to maintain information over prolonged time steps.

Output Gate: Determining the subsequent hidden state based on the cell state and input, the output gate influences the new hidden state. The hidden state is essentially a filtered version of the cell state. Similar to other gates, the output gate's behavior is modulated by a sigmoid activation function, while the cell state undergoes transformation through a tanh function.

2.3.3 Bi-LSTM text classification model

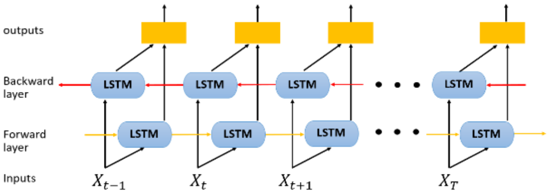

Fig. 3. The structure of Bidirectional – LSTM model.

Bidirectional Long Short-Term Memory (Bi-LSTM) models enhance conventional LSTMs by incorporating a dual-directional processing approach, enabling the model to analyze input data both forward and backward. This bidirectional framework facilitates a comprehensive comprehension of contextual information from preceding and subsequent words in a sequence. Bi-LSTMs demonstrate exceptional efficacy in tasks requiring a nuanced understanding of text, especially when contextual cues are dispersed across various sections of the document. Their heightened capability to capture long-range dependencies arises from the bidirectional processing, making them particularly adept at handling intricate relationships within sequential data.

2.3.4 fastText

fastText, a library developed by Facebook's AI Research (FAIR) lab, is designed for efficient learning of word representations and sentence classification. Its focus is on providing a swift and memory-efficient implementation of algorithms for text classification and representation learning.

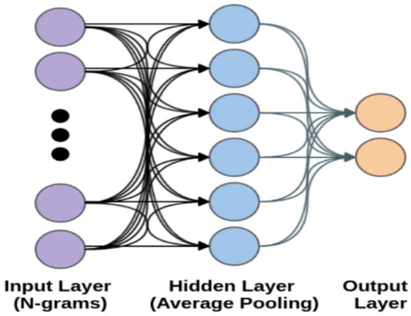

Fig. 4. fastText model structure.

The fastText model utilizes n-gram features for word representation input. To illustrate, consider the sentence "I like apple." When n is set to 2, additional features include the average values of word pairs such as "I like" and "like apple." This incorporation of n-gram features enables fastText to capture word order information in a sentence, leading to more accurate sentence representations. This is a notable advancement compared to traditional bag-of-words models, as it allows for a better understanding of the sequential structure within sentences [16].

2.3.5 WangchanBERTa

WangchanBERTa is currently the largest Thai language model, boasting 106 million parameters for natural language processing tasks. It was trained on the RoBERTa architecture, developed in 2021 through collaboration between VISTEC and PyThaiNLP. This model is an Encoder-only type, trained using Masked Language Modeling (MLM) on diverse public datasets, including Thairath222k, scb-mt-en-th, wongnai-corpus, and the Thai National Corpus, with a total data size of 78.5 GB. WangchanBERTa specifically supports the Thai language and is suitable for applications related to natural language understanding [17].

2.4 Evaluation metrics

In this experiment we used Precision rate, recall rate, F1-score and Hamming loss to evaluate performance of multi label text classification.

Precision is a measure of the accuracy of positive predictions made by a model. It is calculated as the ratio of true positive predictions to the total number of positive predictions made by the model. Precision is particularly useful when the cost of false positives is high, as it reflects the ability of the model to avoid making incorrect positive predictions.

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}.$$

Recall, also known as sensitivity or true positive rate, measures the ability of a model to correctly identify all relevant instances, capturing the ratio of true positives to the total number of actual positive instances. Recall is crucial when it is essential to capture as many positive instances as possible, and the cost of missing a positive instance is high.

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}.$$

The F1-score is a harmonic mean of precision and recall, providing a balanced measure that considers both false positives and false negatives. It is particularly useful in situations where there is an imbalance between positive and negative classes. The F1-score ranges from 0 to 1, where a higher value indicates better model performance. It is beneficial when there is a need to strike a balance between precision and recall, as it penalizes extreme values in either metric.

$$F1\ score = 2 \times \frac{Precision \times Recall}{Precision + Recall}.$$

Hamming Loss is a metric used in multi-label classification tasks to quantify the accuracy of a model in predicting multiple labels for each instance. In multi-label classification, each instance can belong to multiple classes simultaneously. The Hamming Loss measures the fraction of labels that are incorrectly predicted.

$$Hamming\ Loss = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{L}\sum_{j=1}^{L}\left(\hat{y}_{ij} \neq y_{ij}\right).$$

A lower Hamming Loss indicates better performance, with a value of 0 indicating perfect predictions. This metric is useful for evaluating the overall accuracy of a multi-label classification model in handling multiple class assignments for each instance.

3. Methodology

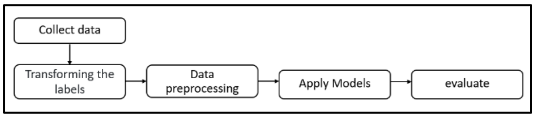

Fig. 5. The flow chart of proposed method.

3.1 Data collecting

The experiment utilized data from reports submitted by residents of Bangkok to the Traffic Fondue platform. These reports, spanning from January 2022 to July 2023, cover diverse issues categorized into 24 distinct categories such as PM2.5, travel, obstructions, safety, cleanliness, traffic, and more. The dataset, obtained from the Traffy Fondue website, consists of 104,970 instances, with 80% (84,000 samples) allocated to the training set and the remaining 20% (20,970 samples) to the test set. Additionally, a test set of 150 instances from the 'Other' label was collected and assigned new labels for further evaluation. The dataset provides a comprehensive basis for analysis and research purposes. More details and access to the dataset can be found at https://share.traffy.in.th/teamchadchart.

3.2 Transforming the labels

The process of transforming labels, specifically converting a label from a list into a binary representation, can be achieved using the MultiLabelBinarizer from the scikit-learn (sklearn) package. This transformation replaces the length of the label list with binary values, where each class is represented as a binary column.

3.3 Data preprocessing

Data preprocessing is a crucial step in data management, especially when dealing with text data that may contain noise or misspellings. In this experiment, data preprocessing was performed to enhance the quality of the dataset for training the model. The PyThaiNLP package was employed as a tool for data cleaning, involving the following key steps:

1. Emoticon Removal: Emoticons were removed from the text data to eliminate non-textual elements that could interfere with analysis.

2. Number Removal: Numeric characters were removed from the text to focus on textual content and reduce numerical noise.

3. Non-Thai Alphabet Removal: Any non-Thai alphabetic characters were eliminated, ensuring that the text data remains consistent with the language of interest.

4. Stop word Removal: Common stop words (such as "ครับ", "ค่ะ", "หรือ", etc.) were removed to improve the relevance of the remaining words in the text.

5. Special Characters and Marks Removal: Special characters and punctuation marks were removed, enhancing the text's readability and consistency.

6. Word Tokenization: Each sentence was further broken down into individual words, a step known as word tokenization.

7. Word Normalization: Incorrect or misspelled words were corrected, ensuring that the text data is more accurate and coherent for further analysis and modeling.

3.4 Apply models

Table 2. The example of multi – label classification problem.

Model	#Param.	Task	Pre-processing
TextCNN	7.46M	Specific training	Token+dict (+ emb + pad)
LSTM	7.53M	Specific training	Token+dict (+ emb + pad)
Bi-LSTM	7.69M	Specific training	Token+dict (+ emb + pad)
CNN+Bi-LSTM	7.60M	Specific training	Token+dict (+ emb + pad)
Wangchan BERTa	110M	Pre -training	Token+dict (+pad)
FastText	6.90M	Specific training	Token+dict (+pad +Ngram)

In the model application phase, we selected six models for evaluation: TextCNN, LSTM, Bi-LSTM, CNN+Bi-LSTM, BERT, and fastText. The first four models incorporated pre-trained embeddings from Facebook, utilizing the fastText algorithm to enhance learning efficiency. For WangchanBERTa, a pre-trained model specifically trained on Thai language data was employed for experimentation. Lastly, the fastText model, which incorporated the N-grams technique, aimed to improve accuracy through the utilization of a different approach.

3.5 Apply models

For the evaluation of a multi-label classification model, we have employed precision, recall, and F1-score, along with the utilization of Hamming Loss to assess the accuracy of predicted labels compared to the ground truth labels for the specified instances. The incorporation of precision, recall, and F1-score allows us to gauge the model's performance across multiple labels, while Hamming Loss provides a measure of how well the predicted labels align with the actual labels for the given instances.

3.6 Results

Table 3 presents the performance metrics of various models in a text classification task. The models evaluated include TextCNN, LSTM, Bi-LSTM, CNN+Bi-LSTM, WangchanBERTa, and fastText. The key metrics reported are precision, recall, F1 score, and Hamming Loss.

Table 3. The example of multi-label classification problem.

Model	Pre-cision	recall	F1-score	Hamming Loss
TextCNN	89.66	83.33	85.66	0.01611
LSTM	90.33	82	85	0.01551
Bi-LSTM	92	86	88	0.01417
CNN+ Bi-LSTM	**92**	**87.33**	**89**	0.01323
Wangchan BERTa	91	86	88	**0.01008**
FastText	92	84.33	88	0.01465

The results indicate that Bi-LSTM achieved the highest precision (92%), recall (86.88%), F1 score (88%), and the lowest Hamming Loss (0.01417) among the individual models. CNN+Bi-LSTM also performed well, achieving a precision of 92%, recall of 87.33%, and F1 score of 89%, with a Hamming Loss of 0.01323.

WangchanBERTa demonstrated competitive performance with a precision of 91%, recall of 86.88%, F1 score of 88%, and the lowest Hamming Loss among all models (0.01008).

fastText exhibited a precision of 92%, recall of 84.33%, F1 score of 88%, and a Hamming Loss of 0.01465.

In summary, the Bi-LSTM and CNN+Bi-LSTM models showed strong overall performance, while WangchanBERTa demonstrated a balance between precision, recall, and Hamming Loss. fastText also performed well but with a slightly higher Hamming Loss compared to other models.

Table 4. Model results of 'others' dataset.

Model	Pre-cision	recall	F1-score	Hamming Loss
TextCNN	35	32	27	0.051
LSTM	37	39	33	0.048
Bi-LSTM	41	35	31	0.048
CNN+ BiLSTM	41	41	35	0.042
Wangchan BERTa	**46**	**49**	**42**	**0.036**
FastText	**46**	33	31	0.046

In the evaluation using the 'Others' dataset Table 4 shows that the models' performance metrics exhibit variations compared to the results from the first dataset.

TextCNN shows a precision of 35%, recall of 32%, F1 score of 27%, and a Hamming Loss of 0.051. LSTM demonstrated slightly improved performance with a precision of 37%, recall of 39%, F1 score of 33%, and a Hamming Loss of 0.048.

Bi-LSTM achieves a precision of 41%, recall of 35%, and F1 score of 31%, with a Hamming Loss of 0.048. The combined model, CNN+Bi-LSTM, exhibited a precision of 41%, recall of 41%, F1 score of 35%, and a Hamming Loss of 0.042.

WangchanBERTa demonstrates the best performance among the evaluated models in the second dataset, with a precision of 46%, recall of 49%, F1 score of 42%, and a Hamming Loss of 0.036.

fastText, on the other hand, showed a precision of 46%, recall of 33%, F1 score of 31%, and a Hamming Loss of 0.046.

WangchanBERTa performed the best on the second dataset, showing higher precision, recall, and F1 score compared to other models. CNN+Bi-LSTM also performed relatively well, while the other models demonstrated varying degrees of effectiveness on this dataset.

In the analysis of the 'Others' dataset, a notable observation emerged: certain data instances proved challenging to classify based solely on textual information. This challenge arose due to the presence of information that inherently requires visual context, necessitating the inclusion of accompanying images for a comprehensive understanding. The nature of the messages suggests that the context of the content and the identification of specific issues therein are reliant on visual elements. Moreover, some messages lacked predefined categories from the platform, further complicating the classification process. Consequently, the model's overall performance exhibited a noticeable decline when compared to its performance on the first dataset.

4. Conclusion

This paper presents deep learning models specifically developed for multi-label text classification, aiming to enhance the efficiency of the Traffy Fondue platform in addressing various issues in Bangkok. The research evaluates various models and emphasizes the importance of incorporating visual elements, particularly for messages lacking predefined categories. The study suggests potential enhancements, such as integrating image data or employing a new categorization method using organizations as labels, to achieve a more comprehensive understanding of complex issues. In summary, the research provides valuable insights for improving text classification models and enhancing context awareness in dealing with diverse problem categories.

References

[1] S. Kim, M. Lee and J. Seok. Multi-label Text Classification of Economic Concepts from Economic News Articles using Natural Language Processing. Thirteenth International Conference on Ubiquitous and Future Networks (ICUFN), Barcelona, Spain, 2022, pp. 417-20,

[2] L. Jia, J. Fan, D. Sun, Q. Gao and Y. Lu. Research on multi-label classification problems based on neural networks and label correlation. 41st Chinese Control Conference (CCC), Hefei, China, 2022, pp. 7298-302.

[3] Q. Zhang, R. Zheng, Z. Zhao, B. Chai and J. Li. A TextCNN Based Approach for Multi-label Text Classification of Power Fault Data. IEEE 5th International Conference on Cloud Computing and Big Data Analytics (ICCCBDA), Chengdu, China, 2020, pp. 179-83.

[4] B. alsukhni. Multi-Label Arabic Text Classification Based On Deep Learning. 12th International Conference on Information and Communication Systems (ICICS), Valencia, Spain, 2021, pp. 475-7.

[5] S. Wang, Y. Yang and X. Meng. Research on Multi-Label Text Classification Based on Multi-Channel CNN and BiLSTM. International Conference on Artificial Intelligence of Things and Crowdsensing (AIoTCs), Nicosia, Cyprus, 2022, pp. 498-503.

[6] D. Kim, J. Koo and U. -M. Kim. EnvBERT: Multi-Label Text Classification for Imbalanced, Noisy Environmental News Data. 15th International Conference on Ubiquitous Information Management and Communication (IMCOM), Seoul, Korea (South), 2021, pp. 1-8.

[7] M. Zang, S. Niu, Y. Gao and X. Chen. Long Text Multi-label Classification," 3rd International Conference on Neural Networks, Information and Communication Engineering (NNICE), Guangzhou, China, 2023, pp. 438-42.

[8] C. Xin and Q. Zhanzhi. Research on Chinese Text Classification Based on WAE and SVM. 3rd International Conference on Natural Language Processing (ICNLP), Beijing, China, 2021, pp. 14-9.

[9] F. Gürcan. Multi-Class Classification of Turkish Texts with Machine Learning Algorithms. 2nd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), Ankara, Turkey, 2018, pp. 1-5.

[10] B. Zhang. News Text Classification Algorithm Based on Machine Learning Technology. International Conference on Education, Network and Information

Technology (ICENIT), Liverpool, United Kingdom, 2022, pp. 182-6.

[11] T. Yao, Z. Zhai and B. Gao. Text Classification Model Based on fastText. IEEE International Conference on Artificial Intelligence and Information Systems (ICAIIS), Dalian, China, 2020, pp. 154-7.

[12] Y. Zhang, C. Liu and W. Liu. A Study on Sentiment Analysis of Movie Reviews based on ALBERT-TextCNN-HAN. 3rd International Symposium on Computer Technology and Information Science (ISCTIS), Chengdu, China, 2023, pp. 758-63.

[13] Y. Lai and L. Zhang. Government affairs message text classification based on RoBerta and TextCNN. 5th International Conference on Communications, Information System and Computer Engineering (CISCE), Guangzhou, China, 2023, pp. 258-62.

[14] M. Gao and J. Li. Chinese short text classification method based on word embedding and Long Short-Term Memory Neural Network. International Conference on Artificial Intelligence, Big Data and Algorithms (CAIBDA), Xi'an, China, 2021, pp. 91-5.

[15] S. Ariwibowo, A. S. Girsang and Diana. Hate Speech Text Classification Using Long Short-Term Memory (LSTM). IEEE International Conference of Computer Science and Information Technology (ICOSNIKOM), Laguboti, North Sumatra, Indonesia, 2022, pp. 1-6.

[16] T. Yao, Z. Zhai and B. Gao. Text Classification Model Based on fastText. IEEE International Conference on Artificial Intelligence and Information Systems (ICAIIS), Dalian, China, 2020, pp. 154-7.

[17] Lalita Lowphansirikul, Charin Polpanumas, Nawat Jantrakulchai and Sarana Nutanong, (2021) Wangchanberta: Pretraining transformer-based thai language models. arXiv preprint arXiv:2101.09635.

[18] Digital Innovation Group for Cities, National Science and Technology Development Agency (NSTDA). Team Chadchart Traffic Dataset. Available: https://share.traffy.in.th/teamchadchart