LegalBERT-th: Development of Legal Q&A Dataset and Automatic Question Tagging

Kannika Wiratchawa
Visual Intelligence Laboratory
Department of Statistics
Faculty of Science
Khon Kaen University
kannikaw@kkumail.com

Tanutcha Khunthong
Visual Intelligence Laboratory
Department of Statistics
Faculty of Science
Khon Kaen University
tanutcha_k@kkumail.com

Thanapong Intharah[†]
Visual Intelligence Laboratory
Department of Statistics
Faculty of Science
Khon Kaen University
thanin@kku.ac.th

***Abstract*—Tagging questions according to their topics is useful for internet forum management. In this paper, we use the Bidirectional Encoder Representations from Transformers (BERT) model to categorize posts from Thai legal internet forums. First, We construct our new legal Q&A dataset by scraping the internet, cleaning the data, and annotating the data. Second, We perform transfer learning to let our model learn about the legal language model in general and then fine-tune the model for the law topic classification task. As a result, we have developed a legal Q&A dataset of 12,695 question/answer pairs and a law topic classification model based on BERT with 92% accuracy. Finally, we build a prototype legal internet forum which equipped with the automatic tagging function, law topic classification, to provide a concrete example of how to apply the model in the real situation.**

***Index Terms*—Bidirectional Encoder Representations from Transformers, Legal Classification, Question Tagging, Legal Dataset, NLP Dataset**

I. Introduction

Large number of people across the globe have very little legal knowledge including knowledge about their right [1]–[4]. In the Information Age, one of the top choices in mind when people have a legal issue is to ask on internet forums. We scope our study in this paper only in Thailand context, where there are several law firms setting up their own internet forums to give advice to public audiences. In the forums, the owners/admins, who are certified lawyers, are actively advising people on their legal issues.

The problem is as the numbers of questions increased, the forums become cluttered. So, the legal advisors have trouble managing the questions and the people who intend to study about relating answered cases have difficult time looking for similar questions.

Hence, we develop a web service which can tag a new input question with a law topic to keep the forum organized. The service is based on a Natural Language Processing (NLP) text classification model. To develop the question tagging service, two steps needed to be carried out.

First, we need to have the dataset especially for the task. However, there does not exist a dataset in Thai language which categorizes informal language that people used in the internet forums into different law topics. Therefore, We have developed the dataset by scraping from legal internet forums, preprocessing the data, and using Latent Dirichlet Allocation (LDA) [5] to assist manual annotation. The process will be described in section III.

Second, We trained a BERT-based model to understand legal language model using transfer learning with the Act of Parliament dataset [6] and our Thai legal Q&A dataset. The model was then trained as a task-specific model to classify law topic of a question posted to the forum. The detail can be found in section IV.

To demonstrate the model in real situation, we implemented the prototype internet forum management system which employed the automatic tagging service. The system architecture is discussed in section V.

Main contributions of this paper are two folds:

- Thai Legal Q&A dataset of 12,695 question/answer pairs, every pair labeled as one of six law topics: Personal Rights, Family, Labor, Contract, and Criminal and *Others*.
- the LegalBERT-th, an NLP model for law topic classification based on the Bidirectional Encoder Representation from Transformers model.

II. Related Work

We discuss in this section about two related topics: text analysis algorithms and legal datasets.

A. Text Analysis Algorithms

Vaswani *et al.,* [7] proposed the Transformer model which introduced multi-head attention. This attention mechanism helps the model understand contexts of words in the input text for the Machine Translation task. The model gave impressive result in English-to-German translation which at that time very challenging. Bidirectional Encoder Representation from Transformers [8] was introduced to solve the Text Classification, Q&A, and Named Entity Recognition tasks by learning from sub-words which decrease size of the vocabulary that the model needs to remember. Moreover, BERT introduced two effective pre-training techniques: masked language modelling (MLM) and next sentence prediction (NSP).

Many works had adopted BERT as one of the top options for Text Classification tasks such as Sentiment Analysis and

[†]corresponding author

camera ready version [2021] submitted to ECTI-CON 18^{th}

Topic Classification. BERT was modified to be able to analyze long documents in [9]. It showed that RoBERT and ToBERT classify customers' opinion more accurate than other NLP models. Moreover, Cui *et al.* [10] used TF-IDF and Naïve Bayes to generate tags for product categories from product posts on Taobao e-commerce platform. They then trained BERT to classify new post into different product categories with up to 90% accuracy. Last but not least, BERT had been illustrated in [11] that it surpassed other algorithm in the German legal document analysis tasks. The tasks discussed the research were classify correct court and level of appeal, predict amount in dispute given a court decision, and calculate semantic similarity between two documents.

In our work, we aim to classify informal questions from the legal internet forum as one of the law topics.

B. Legal Datasets

Holzenberger *et al.,* [12] developed US Tax Law dataset to facilitate Natural Language Processing Q&A tasks. In [13], the Q&A dataset for Chinese Law had been developed. The dataset comprises of 26,365 multiple choice questions in Chinese from the National Judicial Examination of China. The aim is to train a model to predict the answer given the question. Furthermore, [14] developed a dataset of legal agreement which labelled as either fixed term or auto-renewing term to train BERT-based model to help in the intelligent repository system. For Thai language, [15] proposed Wikipedia Corpus in Thai language. The dataset comprises of 5,000 factoids annotated with questions from 7 question words.

In this paper, We develop a Legal Q&A dataset by scraping from Thai legal internet forums. In contrast to other legal dataset the language presented in the question part is informal language while the language presented in the answer is semi-informal. The question/answer pairs is annotated as one of six law topics based on the answer part.

III. Development of the Thai Legal Q&A Dataset

To the best of our knowledge, there does not exist a dataset which meets our criteria: Thai language, informal language used in the internet forums, sentence annotated with law topic(s). Hence, we developed the dataset by following the steps in Fig. 1.

First, we gather legal internet forums and filter the forums by only select the forums which only allow admins who are certified lawyer to answer to the posts. In the first step, we selected six legal in forums out of eight forums. Second, we scraped question/answer from the forums which made up of 21,149 question/answer pairs. Third, the data was then preprocessed by removing HTML tags and stop words. The question/answer pairs were then manually filtered by keeping only the pairs which were in Thai and were about legal advice. At this point, we had 12,695 question/answer pairs. Before passing data to the data analysis and data annotation step, it is important for data in Thai language to perform word tokenization as Thai language does not have space to separate between words. Note that, the stop words removal and the word tokenizer used in this work were from the PyThaiNLP package [16]. In the last step, we used TF-IDF and LDA on the answer parts of the question/answer pairs to come up with unsupervised topics. Please note here that the result of LDA is a set of different topic settings. We selected the setting that maximized the number of main topics and the number of sentences across the main topics. As a result, we ended up with five main topics where the number of documents in the main topics were 7,031 documents. Lastly, the documents from the main topics were manually validated by expert annotators and the five main topics we discovered were Personal Rights, Family, Labor, Contract, and Criminal. The rest of the sentences outside the main topics were categorized as *Others* law topics. We finally assigned the law topics we discovered back to the question/answer pairs, one topic per sentence. The final data distribution is illustrated in Table I and the example of developed dataset is shown in Fig. 2.

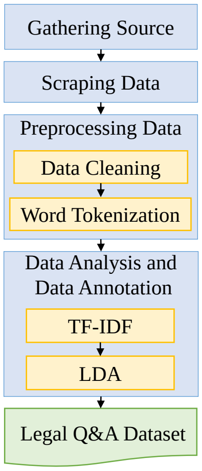


Fig. 1. Development of the Thai legal Q&A dataset.

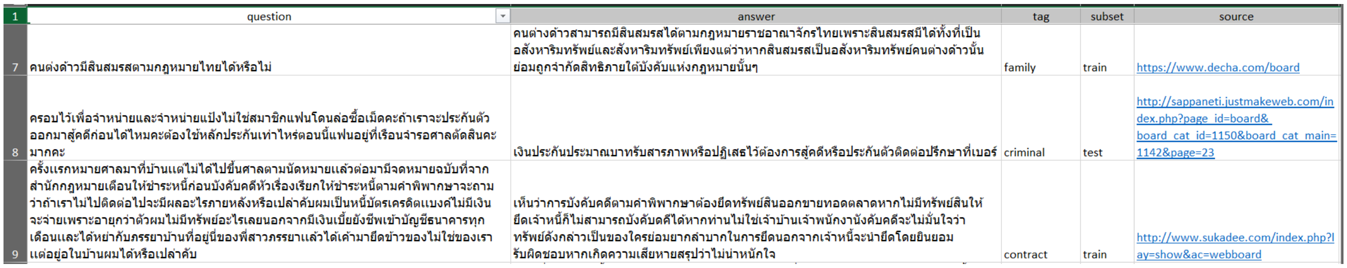
TABLE I
Distribution of law topics in our Thai Legal Q&A dataset.

Law topics	#Question/Answer pairs
Personal rights	382
Family	2,114
Labor	945
Contract	2,534
Criminal	1,056
Others	5,647

IV. Training the LegalBERT-th

From the Thai Legal Q&A dataset, the *Others* topic was used for transfer learning and the five main topics was used for fine-tuning the law topic classification model. The overall steps are demonstrated in Fig. 3.

For transfer learning, we used BERT-th model [17] as our pre-trained weights for Thai language. To make the model learn about law language model, we used the Thai Legal Corpus (Act of Parliament) [6] and the *Other* topic from



1	question	answer	tag	subset	source
7	คนต่งด้าวมีสินสมรสตามกฎหมายไทยได้หรือไม่	คนต่างด้าวสามารถมีสินสมรสได้ตามกฎหมายราชอาณาจักรไทยเพราะสินสมรสมีได้ทั้งที่เป็นอสังหาริมทรัพย์และสังหาริมทรัพย์เพียงแต่ว่าหากสินสมรสเป็นอสังหาริมทรัพย์คนต่างด้าวนั้นย่อมถูกจำกัดสิทธิภายใต้บังคับแห่งกฎหมายนั้นๆ	family	train	https://www.decha.com/board
8	ครอบไว้เพื่อจำหน่ายและจำหน่ายแป้งไม่ใช่สมาชิกแฟนโดนล่อซื้อเม็ดคะถ้าเราจะประกันตัวออกมาสู้คดีก่อนได้ไหมคะต้องใช้หลักประกันเท่าไหร่ตอนนี้แฟนอยู่ที่เรือนจำรอศาลตัดสินคะมากคะ	เงินประกันประมาณบาทรับสารภาพหรือปฏิเสธไว้ต้องการสู้คดีหรือประกันตัวติดต่อปรึกษาที่เบอร์	criminal	test	http://sappaneti.justmakeweb.com/index.php?page_id=board&board_cat_id=1150&board_cat_main=1142&page=23
9	ครั้งแรกหมายศาลมาที่บ้านแต่ไม่ได้ไปขึ้นศาลตามนัดหมายแล้วต่อมามีจดหมายฉบับที่จากสำนักกฎหมายเตือนให้ชำระหนี้ก่อนบังคับคดีหัวเรื่องเรียกให้ชำระหนี้ตามคำพิพากษาจะถามว่าถ้าเราไม่ไปติดต่อไปจะมีผลอะไรภายหลังหรือเปล่าคับผมเป็นหนี้บัตรเครดิตแบงค์ไม่มีเงินจะจ่ายเพราะอายุกว่าตัวผมไม่มีทรัพย์อะไรเลยนอกจากมีเงินเบี้ยยังชีพเข้าบัญชีธนาคารทุกเดือนและได้หย่ากับภรรยาบ้านที่อยู่นี่ของพี่สาวภรรยาแล้วได้เค้ามายึดข้าวของไม่ใช่ของเราแต่อยู่ในบ้านผมได้หรือเปล่าคับ	เห็นว่าการบังคับคดีตามคำพิพากษาต้องยึดทรัพย์สินออกขายทอดตลาดหากไม่มีทรัพย์สินให้ยึดเจ้าหนี้ก็ไม่สามารถบังคับดคีได้หากท่านไม่ใช่เจ้าบ้านเจ้าพนักงานังคับคดีจะไม่มั่นใจว่าทรัพย์ดังกล่าวเป็นของใครย่อมยากลำบากในการยึดนอกจากเจ้าหนี้จะนำยึดโดยยินยอมรับผิดชอบหากเกิดความเสียหายสรุปว่าไม่น่าหนักใจ	contract	train	http://www.sukadee.com/index.php?lay=show&ac=webboard

Fig. 2. Examples of questions, answers, tags, subsets, and sources in our Thai Legal Q&A Dataset.

TABLE II
TRANSFER LEARNING RESULT.

Tasks	Accuracy (%)
Masked Language Modeling (MLM)	37.95
Next Sentence Prediction (NSP)	86.25

our legal Q&A dataset as training data. Note that, we used both question part and answer part of the question/answer pairs for transfer learning. The training tasks in this step are masked language modelling (MLM) and next sentence prediction (NSP) [8]. The result of the transfer learning is showed in Table II.

To fine-tune the model to classify input question as one of the five law topics, we used only question part of the question/answer pairs from the legal Q&A dataset (only pairs from five main topics) as our dataset. The dataset was then divided into 70% training set (4,922 questions) and 30% test set (2,109 questions). We fine-tuned the model from the transfer learning step using Adam Optimizer and set batch size and learning rate to 32 and 5e-5 respectively. The model was trained on NVIDIA Tesla V100 Machine with Tensorflow implementation. The result of the fine-tuned model is showed in Table III.

TABLE III
PERFORMANCE OF LEGALBERT-TH FOR PREDICTING ONE OF FIVE LAW TOPICS FOR THE INPUT QUESTION (WITHOUT DATA AUGMENTATION).

Details	Precision	Recall	F1-Score	Supports
Personal Rights	0.83	0.67	0.74	110
Contract	0.93	0.92	0.92	794
Criminal	0.89	0.9	0.89	297
Family	0.95	0.96	0.96	605
Labor	0.91	0.96	0.93	303
Accuracy			0.92	2109
Macro average	0.9	0.88	0.89	2109
Weighted average	0.92	0.92	0.92	2109

A. Data Augmentation

From Table I, we can see that the data is imbalanced. Hence, in this subsection we used data augmentation techniques to remedy the problem. The data augmentation was performed on the Thai Legal Q&A dataset in the fine-tuning step.

The data augmentation techniques performed in this steps are as follows,

- randomly replace letter(s) with other letters which people usually misspelled.
- randomly remove letter(s).
- randomly add more similar letter after a letter.
- randomly replace word with similar meaning word. (function: most_similar_cosmul() [16])

TABLE IV
PERFORMANCE OF LEGALBERT-TH FOR PREDICTING ONE OF FIVE LAW TOPICS FOR THE INPUT QUESTION (WITH DATA AUGMENTATION).

Details	Precision	Recall	F1-Score	Supports
Personal Rights	0.78	0.73	0.75	110
Contract	0.93	0.91	0.92	794
Criminal	0.90	0.90	0.90	297
Family	0.94	0.97	0.95	605
Labor	0.92	0.95	0.94	303
Accuracy			0.92	2109
Macro average	0.89	0.89	0.89	2109
Weighted average	0.92	0.92	0.92	2109

The goal of data augmentation is to increase the number of questions in each of five law topics to 2,000 questions. We tested different combinations of the data augmentation techniques. The best result when train LegalBERT-th model with augmented data is showed in Table IV. It can be seen that The result is agreed with work from [18], [19] which states that data augmentation did not improve the performance of the model in some learning tasks.

V. LEGALBERT-TH AS A TAGGING SERVICE

In this section, we deployed the model as an automatic question tagging service and connect the service with the real internet forum to demonstrate the direct application of the model and suggest possible setup.

From Fig. 4, The trained LegalBERT-th model from section IV is deployed as RESTful API using Python's Flask package. The RESTful API waits for a HTTP POST request from the internet forum which sends text from the submission form. The text delivered via the HTTP POST request is a question submitted by public users. After the model made prediction, the prediction result is sent back to the forum to use as the tag of the question. This tag is used to filter the question for users and used to organize the post for forum management and data analytics.

VI. DISCUSSION AND CONCLUSION

In this paper, we demonstrate steps toward building text understanding application in real-world context. We started from scraping data from internet forums to building prototype internet forum which deployed with the automatic question tagging. In the model construction step, we chose BERT as

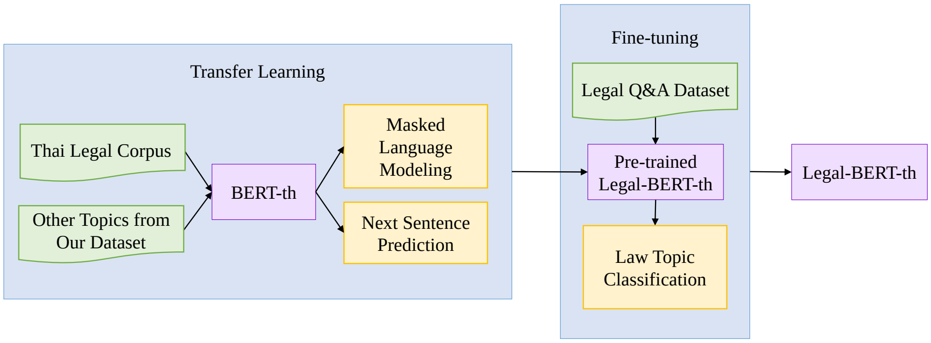


Fig. 3. LegalBERT-th model in based on BERT-th model [17]. The training steps are divided into 2 steps: Transfer Learning and Fine-tuning.

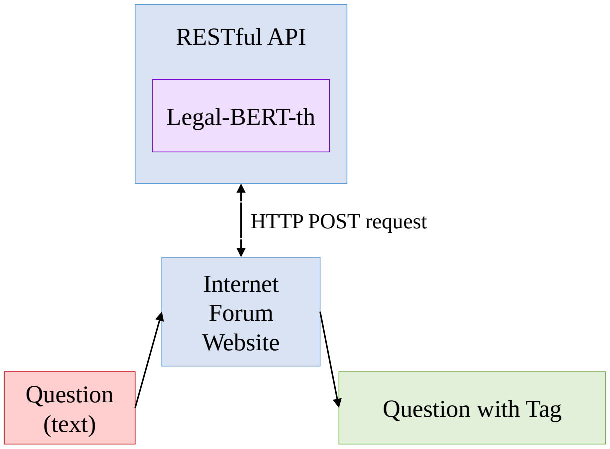


Fig. 4. System architecture of the Legal Internet forum.

our base model and compared the affect of data augmentation technique on the task. The results showed that for both with/without data augmentation the models perform worse on the Personal Rights topic. This can be explained by the dataset distribution which the Personal Rights topic have the smallest number of examples and the data augmentation techniques we deployed could not remedy the data imbalanced problem.

In the future, we aim to develop a multi-label dataset and model that can provide multiple tags to a question. We also intend to extract semantic similarity between two questions for facilitating search. Additionally, the possible further development from this dataset is to build a model for Thai legal question-answering.

Acknowledgment

The authors would like to thank the legal internet forums' owners/admins for kindly providing advice to the legal questions in the forums, Miss Pimnapatr Katesaeng for useful advice on Thai law, and the National Software Contest (NSC-2021) for partially supporting the project.

References

[1] C. Denvir, N. J. Balmer, and P. Pleasence, "When legal rights are not a reality: do individuals know their rights and how can we tell?" *Journal of social welfare and family law*, vol. 35, no. 1, pp. 139–160, 2013.

[2] T. Rejekiningsih, "Law awareness forming strategies to reinforce the principles of social function of land rights within the moral dimension of citizenship," *Procedia-Social and Behavioral Sciences*, vol. 211, pp. 69–74, 2015.

[3] A. Kumar, "National legal literacy mission - an evaluative analysis," *SSRN Electronic Journal*, 03 2013.

[4] K. Alonkorn, "The public awareness of the law," *https://www.senate.go.th/document/Ext21862/21862017_0002.PDF*, 2019.

[5] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *the Journal of machine Learning research*, vol. 3, pp. 993–1022, 2003.

[6] W. Phatthiyaphaibun and A. Suriyawongkul, "Pythainlp: Thai law dataset (act of parliament , jul. 2020," *URL https://github.com/PyThaiNLP/thai-law. Version*.

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2017.

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[9] R. Pappagari, P. Zelasko, J. Villalba, Y. Carmiel, and N. Dehak, "Hierarchical transformers for long document classification," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 838–844.

[10] H. Cui, S. Shao, S. Niu, C. Shi, and L. Zhou, "A classification method for social information of sellers on social network," *EURASIP Journal on Image and Video Processing*, vol. 2021, no. 1, pp. 1–12, 2021.

[11] C. M. Yeung, "Effects of inserting domain vocabulary and fine-tuning bert for german legal language," Master's thesis, University of Twente, 2019.

[12] N. Holzenberger, A. Blair-Stanek, and B. Van Durme, "A dataset for statutory reasoning in tax law entailment and question answering," *arXiv preprint arXiv:2005.05257*, 2020.

[13] H. Zhong, C. Xiao, C. Tu, T. Zhang, Z. Liu, and S. Maosong, "Jec-qa: A legal-domain question answering dataset," in *Proceedings of AAAI*, 2020.

[14] E. Elwany, D. Moore, and G. Oberoi, "Bert goes to law school: Quantifying the competitive advantage of access to large legal corpora in contract understanding," *arXiv preprint arXiv:1911.00473*, 2019.

[15] K. Trakultaweekoon, S. Thaiprayoon, P. Palingoon, and A. Rugchatjaroen, "The first wikipedia questions and factoid answers corpus in the thai language," in *2019 14th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)*. IEEE, 2019, pp. 1–4.

[16] W. Phatthiyaphaibun and K. Chaovavanich, "Pythainlp: a python nlp package for thai, sept. 2016," *URL https://github.com/PyThaiNLP/pythainlp. Version*, vol. 2, no. 6.

[17] T. Team, "Thaikeras:bert pre-training in thai language , dec. 2018," *URL https://github.com/ThAIKeras/bert*.

[18] M. Amjad, G. Sidorov, and A. Zhila, "Data augmentation using machine translation for fake news detection in the urdu language," in *Proceedings of The 12th Language Resources and Evaluation Conference*, 2020, pp. 2537–2542.

[19] R. Jha, C. Lovering, and E. Pavlick, "When does data augmentation help generalization in nlp?" *arXiv preprint arXiv:2004.15012*, 2020.