ROBERTA FOR MULTI-LABEL THAI TEXT CLASSIFICATION

Suwika Plubin
Chiang Mai University, Chiang Mai, Thailand, 50200
Bandhita Plubin
Chiang Mai University, Chiang Mai, Thailand, 50200
Walaithip Bunyatisai
Chiang Mai University, Chiang Mai, Thailand, 50200
Manad Khamkong
Chiang Mai University, Chiang Mai, Thailand, 50200
Thanasak Mouktonglang
Chiang Mai University, Chiang Mai, Thailand, 50200
E-mail: Thanasak.m@cmu.ac.th

Abstract

This study applies the RoBERTa model for multi-label classification of Thai-language customer reviews in the banking sector, utilizing a dataset of 24,500 reviews labeled into multiple categories. The objective is to assess RoBERTa's ability to handle complex linguistic structures and imbalanced data while categorizing reviews into multiple relevant labels. RoBERTa's transformer-based architecture, with its self-attention mechanism, is highly effective in capturing the contextual meaning of Thai text, a language known for its unique challenges, such as lack of spaces between words and tonal variations. The model demonstrated strong performance, achieving a macro average precision of 0.83, an F1-score of 0.71, and a Hamming Loss of 0.083. SMOTE was employed to improve recall in underrepresented categories, enhancing the overall performance balance. The results highlight RoBERTa's effectiveness in Thai-language multi-label text classification, showcasing its capability to manage imbalanced data and deliver accurate, context-aware predictions across multiple categories.

Keywords: RoBERTa, Multi-label classification, Transformer model

Introduction

Text classification is a critical task in natural language processing (NLP) that aims to organize and categorize large volumes of textual data into predefined categories. It is commonly used in applications such as sentiment analysis, spam detection, and document classification (Lv et al., 2024). However, classifying text into multiple categories simultaneously, known as multi-label classification, introduces additional complexity, especially when dealing with underrepresented languages like Thai.

Multi-label text classification presents unique challenges, particularly in Thai language contexts where tokenization and linguistic nuances differ significantly from widely studied languages such as English. Previous approaches to text classification often relied on shallow machine learning models like support vector machines (SVM), k-nearest neighbors (KNN), and naive Bayes (NB) (Liu et al., 2019). While these models perform well for simpler tasks, they require extensive feature engineering and fail to capture complex relationships in text, such as word order and semantic meaning (Tan et al., 2023).

In recent years, transformer-based models have revolutionized text classification tasks. Models like BERT (Bidirectional Encoder Representations from Transformers) have demonstrated superior performance by learning deep contextual relationships between words in a sentence, enabling better semantic understanding (Devlin et al., 2018). However, RoBERTa (Robustly Optimized BERT Pretraining Approach), a variant of BERT, has been shown to outperform BERT by adjusting key training strategies such as removing the next-sentence prediction task and increasing batch sizes (Liu et al., 2019). RoBERTa has become a preferred choice for text classification tasks because of its ability to handle large datasets and complex relationships within the text.

Despite the growing use of RoBERTa in various NLP tasks, its application in Thai multi-label text classification has been relatively unexplored. The Thai language poses significant challenges for NLP models due to its lack of word boundaries, complex grammar, and tonal variations (Tan et al., 2022). Additionally, many Thai text datasets suffer from class imbalance, where some categories have significantly fewer samples than others. This imbalance can lead to poor classification performance, especially in minority classes. Recent studies have addressed these challenges by combining RoBERTa with techniques such as oversampling and the Synthetic Minority Over-sampling Technique (SMOTE) to improve performance on imbalanced datasets (Lv et al., 2024).

This research aims to evaluate the effectiveness of RoBERTa for multi-label Thai text classification in banking customer reviews. By utilizing RoBERTa's deep contextual understanding, we aim to overcome the challenges of imbalanced data and the complex linguistic features of Thai. The study assesses RoBERTa's performance through metrics such as precision, recall, F1-score, accuracy, and Hamming loss, providing a comprehensive analysis of its suitability for multi-label Thai text classification.

Research Objectives

1) Evaluate the effectiveness of RoBERTa in Thai multi-label text classification.
2) Address the issue of imbalanced data in multi-label classification tasks.

Literature Review and Related Work

Text classification has evolved significantly with the development of machine learning and deep learning models. Early approaches, such as Naïve Bayes (NB), Support Vector Machines (SVM), and Decision Trees (DT), were highly popular due to their simplicity and effectiveness in handling small datasets. However, these traditional models faced challenges when applied to multi-label classification tasks, as they relied heavily on manual feature engineering methods like bag-of-words and TF-IDF, which failed to capture the deeper contextual relationships in text. Sonawane et al. (2023) suggested that these traditional methods were inadequate for addressing the complexities of multi-label classification tasks, particularly when dealing with large datasets or more intricate forms of textual data. Similarly, Patwardhan et al. (2023) highlighted that traditional model, such as SVMs and logistic regression, often lack the capability to handle the sequential and contextual nature of language, making them less effective for tasks requiring deep contextual understanding. In contrast, transformer models like RoBERTa, with their self-attention mechanisms, excel at capturing relationships across entire sequences, thereby overcoming these limitations (Vaswani et al., 2017).

To overcome these limitations, deep learning models such as Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) were introduced. These models allowed for automatic feature extraction from raw text data, enabling better performance in more complex classification tasks. Tan et al. (2023) demonstrated that CNNs and RNNs could improve the modeling of sequential dependencies in text, particularly with the use of Long Short-Term Memory (LSTM) networks. However, even with these improvements, challenges like the vanishing gradient problem persisted in RNNs, especially when dealing with longer text sequences. Cheruku et al. (2023) proposed a solution to this problem by integrating RoBERTa, a transformer-based model, with RNNs to improve sentiment classification on social media data. Their work emphasized the potential of combining transformers with RNNs to capture both contextual embeddings and sequential information.

The introduction of transformer models, particularly BERT (Bidirectional Encoder Representations from Transformers), marked a major breakthrough in natural language processing. Devlin et al. (2018) noted that BERT's ability to capture bidirectional context allowed it to outperform previous models in tasks like sentiment analysis and named entity recognition. RoBERTa, an optimized version of BERT, further improved on these capabilities by refining pre-training techniques, such as increasing batch sizes and removing the next-sentence prediction task. Liu et al. (2019) showed that these enhancements led to RoBERTa's superior performance across a wide range of NLP tasks, including text classification and sentiment analysis.

In recent years, RoBERTa has been widely adopted for multi-label classification tasks. Sonawane et al. (2023) demonstrated RoBERTa's effectiveness in their **VaxVerdict** model, which classified COVID-19 vaccine-related tweets into categories like conspiracy theories and vaccine side effects. They achieved a Jaccard score of 0.67 and a macro-F1 score of 0.65, proving RoBERTa's robustness in handling multi-label classification problems. Additionally, Cheruku et al. (2023) proposed a modified RoBERTa model integrated with RNNs for Twitter sentiment classification, showing significant improvements in performance metrics such as accuracy, precision, recall, and F1-score. Their experiments highlighted that combining RoBERTa with RNN architectures enhances classification performance, particularly when dealing with social media data.

RoBERTa-based models have also been effectively used to address the issue of imbalanced datasets in multi-label classification. Zhang et al. (2023) showed that combining RoBERTa with the Synthetic Minority Over-sampling Technique (SMOTE) improved the model's ability to handle minority classes in imbalanced datasets. Similarly, Nguyen et al. (2023) explored the use of RoBERTa with SMOTE to improve classification performance, particularly in cases where certain classes were underrepresented in the dataset. These studies underscore the versatility of RoBERTa in various NLP tasks and its ability to adapt to specific challenges like class imbalance.

Despite these advancements, the application of RoBERTa to Thai-language text classification, particularly in multi-label tasks, remains underexplored. Thai presents unique challenges due to its lack of word boundaries and tonal variations, which complicate tokenization and contextual understanding. This research seeks to address these challenges by applying RoBERTa to the classification of Thai-language banking reviews, aiming to improve both the linguistic handling and classification accuracy in imbalanced datasets.

Scope of the Research

This research focuses on applying RoBERTa, a transformer-based language model, to the multi-label classification of Thai-language banking customer reviews. The study covers key aspects of multi-label text classification, handling imbalanced datasets, and model

fine-tuning, incorporating theoretical foundations and mathematical formulations to guide the methodology.

1) Multi-Label Classification of Thai-Language Text

In multi-label classification, each instance (a customer review) may belong to more than one category. The model assigns multiple labels by predicting probabilities for each label. Mathematically, the problem is formulated as follows:

Let x represent the input text and $y = [y\_1, y\_2, y\_3, \cdots, y\_n]$ represent the set of predicted labels, where $y\_i \in \{0,1\}$ for each category i. The predicted probability for each label is given by:

$$P(y\_i \mid x) = \sigma(W \cdot h(x) + b)$$

where:

- $h(x)$ is the embedding generated by RoBERTa for the input text x,
- W is the weight matrix,
- b is the bias term, and
- σ is the sigmoid activation function, which ensures that each label is predicted independently.

RoBERTa's architecture is based on the transformer model, and its use of the self-attention mechanism allows each token to attend to all others in the input sequence. The attention mechanism is mathematically represented as:

$$Attention(Q, K, V) = softmax((QK\text{^}T)/\sqrt{(d\_k)})V$$

where Q, K, V are query, key, and value matrices, respectively, and d_k is the dimensionality of the keys (Liu et al., 2019).

2) Handling Imbalanced Data

Imbalanced data is a common issue in real-world datasets, including the Thai banking reviews analyzed in this study. SMOTE (Synthetic Minority Over-sampling Technique) is employed to address this issue by generating synthetic samples for underrepresented classes (Zhang et al., 2023). SMOTE works by generating a synthetic data point x_new between two minority class data points x_i and x_j.:

$$x\_new = x\_i + \lambda\,(x\_j - x\_i), \quad \lambda \in [0,1]$$

where λ is a random value that ensures the new point lies along the line segment between x_i and x_j. This technique helps balance the dataset and improve the model's ability to classify minority classes.

3) RoBERTa Model Fine-Tuning and Optimization

RoBERTa builds upon the BERT model by improving its training dynamics. One key change is the removal of the Next Sentence Prediction (NSP) task, which RoBERTa eliminates to improve performance. Instead, RoBERTa uses dynamic masking during the pre-training process, allowing the model to see different masked tokens across training epochs (Liu et al., 2019). The Masked Language Model (MLM) objective is defined as:

$$L\_MLM = -\sum\_(i = 1)\text{^}n\ logP(x\_i \mid x\_(\backslash i)\}$$

where x_i is the masked token, and $x\_(\backslash i)$ represents the remaining tokens in the sequence.

In the fine-tuning phase, Binary Cross-Entropy Loss is used for multi-label classification:

$$L = -\sum_{i=1}^{n} [y_i \log \quad \log(p_i) + (1 - y_i) \log \quad \log(1 - p_i)]$$

where:

- y_i is the true label for category i
- p_i is the predicted probability for category i.

RoBERTa's fine-tuning is optimized using the Adam Optimizer, which adapts the learning rate based on the gradient's first and second moments. The update rule for the weights is given by:

$$\theta\_t = \theta\_(t-1) - \eta\, m\_t/(\sqrt{(v\_t)} + \varepsilon)$$

where η is the learning rate, m_t is the first moment (mean), and v_t is the second moment (uncentered variance) of the gradient (Liu et al., 2019).

4) Evaluation of Transformer-Based Models

The performance of RoBERTa is compared with traditional machine learning models, such as Support Vector Machines (SVM) and deep learning models like BiLSTM. SVM constructs a decision boundary between classes by maximizing the margin between support vectors:

$$w \cdot x + b = 0$$

where w is the weight vector, and b is the bias term.

RoBERTa, with its transformer-based architecture, excels in capturing long-range dependencies and contextual relationships, making it superior to models like SVM in multi-label classification tasks (Cheruku et al., 2023)

5) Performance Metrics

The model's performance is evaluated using key metrics: Precision, which measures the proportion of correctly predicted labels among all predictions; Recall, which assesses the proportion of correctly predicted labels among all actual labels; F1 Score, a harmonic mean of precision and recall that balances the trade-off between the two; and Hamming Loss, which calculates the fraction of incorrectly predicted labels, including both false positives and false negatives, relative to the total number of labels. (Zhang et al., 2023).

Research Methodology

This section outlines the methodology used for applying RoBERTa in multi-label classification of Thai-language banking customer reviews. The methodology is structured into several key phases: dataset preparation, preprocessing, model architecture, training and evaluation, and performance measurement.

1) Dataset Preparation

The dataset consists of 24,500 Thai-language customer reviews from sources like Facebook, X (formerly Twitter), and Pantip, categorized into eight labels: Accessibility, Chatbot, Facility and Support, Image, Product and Services, Timing, Staff, and Other. As a multi-label classification task, each review can belong to multiple categories. The dataset is imbalanced, with certain categories like Chatbot and Facility being underrepresented, resulting in a total of 67,870 labeled sentences.

2) Data Preprocessing

To prepare the dataset for the RoBERTa model, several preprocessing steps were applied. Since Thai text lacks spaces between words, tokenization is a complex task. A Thai-specific tokenizer from the *PyThaiNLP* library was used to segment the text into tokens, enabling the model to process the input effectively. Common stopwords in Thai, such as "ที่" (that), "ของ" (of), and "และ" (and), were removed to reduce noise and ensure that only relevant words were used for classification. To handle the class imbalance in the dataset, SMOTE was employed, generating synthetic samples for the underrepresented categories and balancing the dataset. This step helped the model better classify all categories, particularly those with fewer instances. Finally, pre-trained RoBERTa embeddings were used to convert the tokenized text into dense vectors, capturing the contextual meaning of words and sentences in Thai, ensuring that the model could interpret the text accurately and efficiently.

3) Model Architecture

The core architecture of the model is built on RoBERTa, a transformer-based model that uses self-attention to capture token relationships. The RoBERTa encoder generates contextual embeddings that capture both syntactic and semantic meanings. A custom classification layer with a sigmoid activation function is added to predict the probability of each label independently for the multi-label task. The model is trained using Binary Cross-Entropy Loss, suitable for treating each label separately in multi-label classification.

4) Training and Evaluation

The RoBERTa model is fine-tuned for Thai text classification through a structured process. It utilizes the Adam optimizer with a learning rate of 1e-5 and a batch size of 32, trained for 100 epochs with early stopping to prevent overfitting. Dropout regularization is also applied during classification. A stratified 10-fold cross-validation approach is employed to maintain balanced category distribution, providing a comprehensive assessment of model performance. Additionally, hyperparameter tuning is conducted using grid search to optimize key parameters, ensuring the model operates with optimal settings for accurate classification.

5) Performance Metrics

The performance of the RoBERTa model is evaluated using precision, recall, and F1-score for each category. The macro-averaged F1-score is used to assess overall performance, balancing precision and recall across all categories.

Research Results

The performance of the RoBERTa model for multi-label classification of Thai-language banking reviews is summarized in Table 1, which presents key metrics such as precision, recall, F1-score, and Hamming Loss for each category. The overall performance is evaluated using macro averages and Hamming Loss, providing a comprehensive overview of the model's effectiveness across all categories.

Table 1: Performance metrics of RoBERTa on multi-label classification of Thai-language banking customer reviews.

Category	Precision	Recall	F1-Score	Hamming Loss
Accessibility	0.87	0.75	0.80	
Chatbot	0.65	0.50	0.57	
Facility and Supporter	0.70	0.55	0.61	
Image	0.85	0.60	0.70	
Other	0.83	0.68	0.75	0. 08318
Product and Services	0.86	0.65	0.74	
Staff	0.88	0.65	0.75	
Timing	0.80	0.65	0.72	
macro avg	0.83	0.63	0.71	

The RoBERTa model demonstrated effective performance in multi-label classification of Thai-language banking reviews, achieving high precision in categories like *Accessibility* (precision: 0.87, F1-score: 0.80) and *Product and Services* (precision: 0.86, F1-score: 0.74). However, it struggled with underrepresented categories, such as *Chatbot* (precision: 0.65, F1-score: 0.57) and *Facility and Supporter* (precision: 0.70, F1-score: 0.61). The overall macro average precision was 0.83, while the macro average recall was 0.63, indicating challenges in retrieving relevant instances from less populated categories. The macro average F1-score of 0.71 reflects a balanced performance, and the Hamming Loss of 0.08318 signifies relatively few incorrect predictions across all categories. These findings confirm RoBERTa's capability in handling complex text classification tasks, especially with balanced datasets.

Conclusion and Discussion

This study applied the RoBERTa model for multi-label classification of Thai-language banking reviews, effectively addressing issues of imbalanced data and the linguistic challenges unique to Thai. The model performed well in well-represented categories, achieving an F1-score of 0.80 for Accessibility and 0.7 for Product and Services. However, it faced challenges in underrepresented categories like Chatbot and Facility and Supporter, where the recall was notably lower. RoBERTa's transformer-based architecture, with its self-attention mechanism, excelled in capturing contextual meaning, achieving a macro average precision of 0.83 and an F1-score of 0.71. This performance surpasses that of traditional models like RNNs and CNNs (Devlin et al., 2018; Liu et al., 2019). Unlike prior research that combined RoBERTa with RNNs (Cheruku et al., 2023), our standalone RoBERTa model demonstrated robust performance, particularly with sufficient data. The implementation of SMOTE enhanced recall in minority classes, although further improvements are necessary for underrepresented categories (Zhang et al., 2023). Compared to traditional models such as SVM and Naïve Bayes, RoBERTa significantly improved precision and contextual understanding, eliminating the need for manual feature engineering commonly required by traditional methods. This advantage arises from RoBERTa's deep transformer architecture, which effectively captures complex relationships in text sequences (Devlin et al., 2018; Liu et al., 2019). Furthermore, RoBERTa's adaptability to non-English languages was confirmed in this Thai-language context, supporting findings from previous studies (Sonawane et al., 2023). Nonetheless, challenges remain, particularly with imbalanced data and distinguishing between closely related categories such as Product and Services and Facility and Supporter. Future work should explore advanced data

augmentation techniques and hierarchical attention mechanisms (Tan et al., 2023) to enhance performance in these underrepresented categories. Overall, RoBERTa is a powerful tool for Thai-language multi-label classification, especially in categories with ample data.

References

Cheruku, R., et al. (2023). Sentiment classification with modified RoBERTa and recurrent neural networks. *Multimedia Tools and Applications*. https://doi.org/10.1007/s11042-023-16833-5

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint* arXiv:1810.04805.

Liu, Y., et al. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint* arXiv:1907.11692.

Lv, S., et al. (2024). RB-GAT: A text classification model based on RoBERTa-BiGRU with Graph ATtention Network. *Sensors*, 24(11), 3365. https://doi.org/10.3390/s24113365

Nguyen, H. T., & Do, P. H. (2023). Multi-label text classification using transformers. *CEUR Workshop Proceedings*, 3681, 105-115. https://ceur-ws.org/Vol-3681/T1-8.pdf

Patwardhan, N., et al. (2023). Transformers in the real world: A survey on NLP applications. *Applied Intelligence*, 53, 1234–1250. https://doi.org/10.1007/s10489-023-04329-1

Zhang, X., & Li, Y. (2023). Multi-label text classification with transformers and SMOTE. *Journal of Big Data*, 10, 1-12. https://doi.org/10.1186/s40537-023-00691-4

Sonawane, S., et al. (2023). VaxVerdict: A RoBERTa-based multi-label tweet classifier. Presented at the FIRE 2023 AISoMe Track Task. SCTR's Pune Institute of Computer Technology, Dhankawadi, Pune.

Tan, K. L., Lee, C. P., & Lim, K. M. (2022). Improving sentiment classification using a RoBERTa-based hybrid model. *Frontiers in Artificial Intelligence*. https://www.frontiersin.org/articles/10.3389/frai.2023.1012123/full

Tan, K. L., et al. (2023). Improving sentiment classification using a RoBERTa-based hybrid model. *Frontiers in Artificial Intelligence*. https://www.frontiersin.org/articles/10.3389/frai.2023.1012123/full

Tan, K. L., et al. (2023). RoBERTa-GRU: A hybrid deep learning model for enhanced sentiment analysis. *Applied Sciences*, 13(6), 3915. https://doi.org/10.3390/app13063915

Vaswani, A., et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5998–6008. https://doi.org/10.48550/arXiv.1706.03762