Journal of Information Systems Engineering and Management
2025, 10(56s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Hierarchical Label-Wise Attention for Imbalanced Multi-Label Thai Text Classification

Suwika Plubin[1], Bandhita Plubin[1], Walaithip Bunyatisai[1], Thanasak Mouktonglang[2], Manad Khamkong[1]*
[1]*Department of Statistics, Faculty of Science, Chiang Mai University, Chiang Mai, Thailand, 50200*
[2]*Department of Mathematics, Faculty of Science, Chiang Mai University, Chiang Mai, Thailand, 50200*
**Corresponding Author: Manad Khamkong, manad.khamkong@cmu.ac.th*
This work was previously presented as an abstract at IIARP International Conference (Vol. 3, No. 3, March 2025)

ARTICLE INFO

Received: 30 Dec 2024

Revised: 12 Feb 2025

Accepted: 26 Feb 2025

ABSTRACT

Introduction: Multi-label text classification, in which each instance can belong to several categories, is vital for applications like sentiment analysis, document classification, and mining customer feedback. In real-world situations, customer feedback frequently covers various subjects at once, highlighting the need to correctly pinpoint all pertinent elements. In practice, real-world datasets often experience significant class imbalance, leading models to prioritize majority classes (e.g., Product & Services, Accessibility) while performing poorly on minority classes (e.g., Chatbot, Facility & Supporter). These difficulties are increased when handling Thai text, which does not have clear word boundaries and includes intricate syntax, making tokenization and contextual comprehension more challenging. Presented at the IIARP International Conference Abstract Proceedings Volume (Vol. 3, No. 3 | March 2025), this study presents an innovative approach designed to address these challenges.

Objectives: This research presents the Hierarchical Label-Wise Attention Transformer (HiLAT) aimed at addressing imbalanced multi-label classification for reviews from Thai banking customers. We aim to enhance overall predictive accuracy and F1-scores per category—particularly for less represented labels—by tailoring attention mechanisms and adjusting loss weighting to the specific traits of Thai.

Methods: We gathered 24,500 customer reviews in Thai (67,870 labeled sentences) from social media, which were manually categorized into eight groups: Accessibility, Chatbot, Facility & Support, Image, Other, Product & Services, Staff, and Timing. To address class imbalance, we utilized a class-weighted loss function in training, enhancing the impact of minority labels. The HiLAT framework includes two attention layers: (1) attention at the sentence level to pinpoint the most pertinent text segments for each label; and (2) token attention by label to concentrate on the most significant tokens within those segments. Pre-trained Thai word vectors were integrated to enhance semantic representations. We evaluated the model's performance by employing macro-averaged precision for label retrieval accuracy, macro-averaged recall for completeness, macro-averaged F1-score for overall balance, and Hamming Loss to measure misclassification rates.

Results: HiLAT achieved a macro-average F1-score of 0.597 with a Hamming Loss of 0.233. It performed best on well-represented labels—Product & Services (F1 0.732), Accessibility (0.725), and Staff (0.627)—and moderately on mid-frequency labels Timing (0.599) and Other (0.547). Performance on low-frequency categories remained lower—Chatbot (0.396), Facility & Supporter (0.415), and Image (0.490)—highlighting opportunities for further enhancement.

Conclusions: By integrating hierarchical attention and class weighting, HiLAT effectively addresses the dual challenges of multi-label prediction and severe class imbalance in Thai text. While further enhancements—such as advanced resampling or data augmentation—may improve performance on minority categories, the strong macro-average metrics validate HiLAT's applicability for nuanced feedback analysis in banking and other sectors.

Keywords: Hierarchical Label-Wise Attention Transformer, Multi-label Text Classification, Transformer, Customer Feedback Analysis

Journal of Information Systems Engineering and Management
2025, 10(56s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

INTRODUCTION

Multi-label text classification, where each text instance may belong to several categories, is a challenging yet crucial task in various applications, such as sentiment analysis, document classification, and customer feedback evaluation. In these tasks, the model has to not only classify the content correctly but also handle the concurrent prediction of several labels. The difficulty is heightened when working with imbalanced datasets, where some categories are markedly less represented than others. Imbalanced datasets present a well-recognized issue in machine learning (He & Garcia, 2009), since models often favor majority classes, resulting in reduced performance for minority categories. Tackling this disparity is essential for guaranteeing both equity and precision across all categories in multi-label classification.

In recent years, deep learning models, particularly those utilizing attention mechanisms, have demonstrated remarkable effectiveness in multi-label text classification. Attention-based models, like Transformers (Vaswani et al., 2017), have transformed natural language processing (NLP) by allowing models to focus specifically on the most relevant segments of the input sequence. These models have shown significant progress over traditional methods, especially in handling tasks that require understanding complex connections between words or phrases in written content. Despite these advancements, utilizing such models for imbalanced multi-label classification tasks, especially in languages with unique structures like Thai, is still insufficiently explored.

The Thai language presents additional challenges due to its absence of clear word boundaries, which complicates tokenization and feature extraction significantly more than in languages such as English (Chantree et al., 2021). Moreover, customer feedback often includes various perspectives: remarks on product quality, service, and usability can all be contained in one message, necessitating models to allocate multiple labels simultaneously. These difficulties are intensified by the reality that certain feedback categories emerge significantly less frequently than others, which makes accurate prediction of uncommon labels particularly difficult in practical datasets.

In this research, we introduce the Hierarchical Label-Wise Attention Transformer (HiLAT), a model designed for the imbalanced, multi-label classification of Thai text. HiLAT utilizes a two-phase attention approach: initially, it implements hierarchical attention to identify the most relevant sentence parts for each label, followed by applying label-specific token-level attention to focus on the crucial words within those parts. This multilevel design improves the model's capacity to understand relationships between words and sentences, even in intricate or lengthy evaluations. By assigning separate attention streams to each label, HiLAT successfully focuses on the pertinent context required to address significant class imbalance.

Our dataset consists of Thai banking customer reviews, which show a significant imbalance across its eight categories: labels like Accessibility, Product & Services, and Staff are abundant, while Chatbot and Facility & Supporter seldom emerge. This imbalanced distribution inhibits the performance of traditional classifiers, which frequently focus on the primary labels while neglecting the rare ones (Buda et al., 2018). Our goal is to create a model that can balance performance between frequent and rare classes by directly addressing the natural imbalance in the dataset.

We evaluate the effectiveness of HiLAT through essential multi-label metrics: the F1-score, which balances precision and recall for every label, and Hamming Loss, which measures the total rate of incorrect label assignments across all samples. Hamming Loss is particularly insightful in multi-label scenarios as it measures the ratio of incorrectly predicted labels to the total potential label assignments (Tsoumakas & Katakis, 2007).

OBJECTIVES

Design a Specialized Architecture: Introduce the Hierarchical Label-Wise Attention Transformer (HiLAT), which employs hierarchical attention to independently focus on text spans and tokens relevant to each label.

Mitigate Class Imbalance: Implement and evaluate class-weighted loss functions to improve model performance on minority labels without sacrificing accuracy on majority classes.

Leverage Thai-Specific Embeddings: Integrate pre-trained Thai word embeddings to enhance the model's ability to capture linguistic nuances and complex structures in Thai text.

Comprehensive Evaluation: Assess HiLAT on a Thai banking customer feedback dataset using standard multi-label metrics (F1-score, Hamming Loss) and compare against traditional baselines.

Journal of Information Systems Engineering and Management
2025, 10(56s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Advance Research in Thai NLP: Contribute insights to the limited body of work on imbalanced multi-label classification in Thai, highlighting challenges and effective strategies for this underrepresented language.

METHODS

In this work, we introduce the HiLAT framework to tackle imbalanced multi-label classification of Thai banking customer feedback. Our approach follows a clear sequence: data acquisition, preprocessing, feature encoding, and finally, deploying the HiLAT model to generate label predictions. The overall process is depicted in Figure 1.

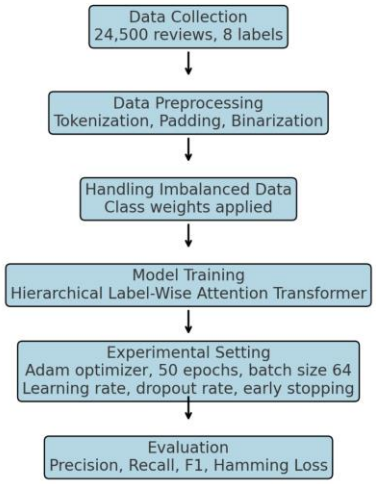

Figure 1: illustrates the workflow of the proposed methodology.

Data Collection

We compiled 24,500 Thai banking customer reviews from platforms including Facebook, X (formerly Twitter), and Pantip. Three expert linguists manually assigned each review to one or more of eight categories: Accessibility, Chatbot, Facility & Supporter, Image, Other, Product & Services, Staff, and Timing. Figure 2 visualizes the number of labeled sentences per category, highlighting the pronounced imbalance among these classes.

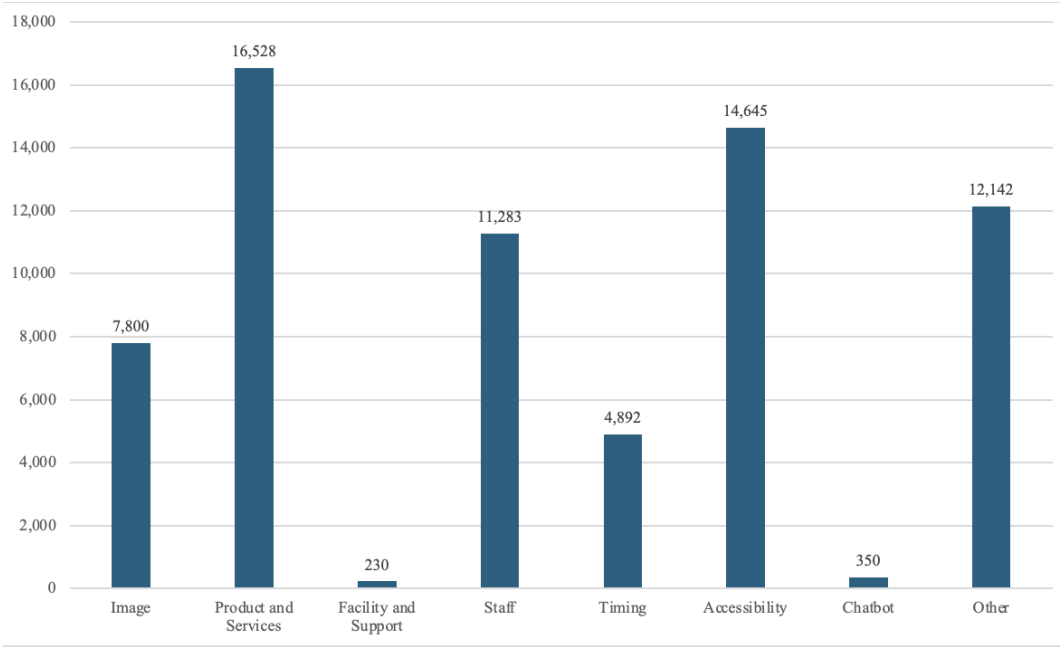

Figure 1. Distribution of Labeled Sentences Across the Eight Categories

Journal of Information Systems Engineering and Management
2025, 10(56s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Data Pre-Processing

Data preprocessing is an essential phase in readying unprocessed text data for deep learning models, particularly in the realm of multi-label text classification. In this research, we utilized a dataset containing 24,500 customer reviews in Thai, collected from social media sites including Facebook, X (previously known as Twitter), and Pantip. Linguists manually classified the reviews into eight categories. Due to the characteristics of the dataset—multi-label, imbalanced, and in Thai—various preprocessing steps were necessary to confirm that the data was formatted appropriately for training the model. These actions consist of:

Text Tokenization

Since Thai writing lacks spaces to distinguish words, dividing sentences into coherent units is more difficult than in English. Tokenization, which transforms raw text into separate words or subwords, is an essential step in preprocessing. We utilized a tokenizer specifically tailored for Thai, which divides each review into the correct word segments. This conversion is crucial since our model functions using numerical representations of tokens instead of unprocessed strings.

Padding Sequences

After tokenization, the lengths of the resulting sequences (reviews) vary, which poses a challenge for models that require fixed-length input. To address this, the pad_sequences function was used to pad or truncate sequences to a uniform length (max_len). This ensures that each review has the same number of tokens, making it possible to batch and process them efficiently in the deep learning model.

Label Binarization

As the task involves multi-label classification, every review may belong to multiple categories. The MultiLabelBinarizer was used to transform the labels for each review into a binary matrix. Every column in this matrix corresponds to one of the eight categories, and the values in the matrix are binary (0 or 1), signifying the presence or lack of each label. This adjustment is essential for managing multi-label data within a supervised learning context.

Handling Imbalanced Data

Some categories such as Chatbot and Facility & Supporter—were severely underrepresented compared to others like Product & Services and Accessibility. To counter this, we introduced class weighting during training, boosting the loss contribution of rarer labels so the model learns to recognize them more reliably without degrading performance on common categories. Additionally, we applied the SMOTE algorithm to synthetically generate minority instances, resulting in a more uniform class distribution across all eight labels. The updated label counts are visualized in Figure 3.

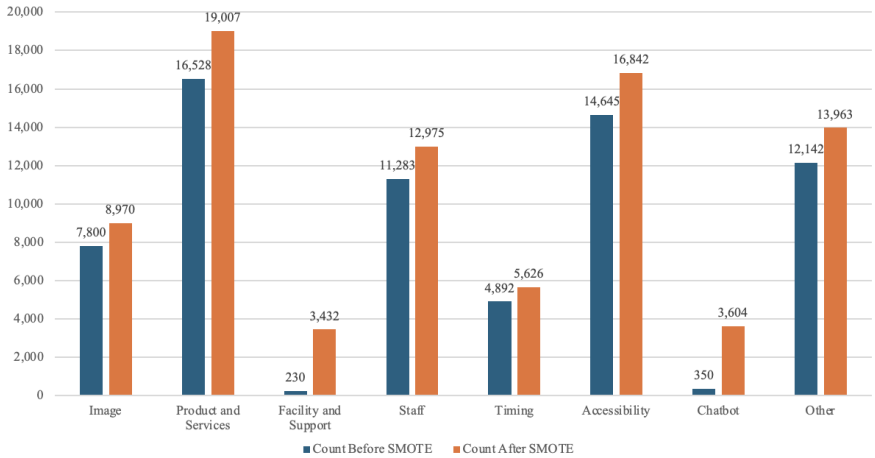

Figure 3. Label Distribution Before and After SMOTE Application

Journal of Information Systems Engineering and Management
2025, 10(56s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Experimental Setting

In the model application phase, we focused on the HiLAT for evaluation. This model was selected due to its ability to capture dependencies in multi-label text classification, where the hierarchical attention mechanism can focus on the most relevant parts of customer feedback for each label. The model incorporates pre-trained embeddings for the Thai language to ensure that the nuances of Thai text are adequately represented.

The HiLAT follows a structured architecture that enables it to learn from the input text and focus on relevant words and sentences for each individual label. The architecture begins with an **Embedding Layer**, where pre-trained Thai word embeddings transform the input sequences into dense vector representations, capturing the semantic relationships between words. Following this, the model's core component, the **Multi-Head Attention** mechanism, allows the model to attend to different parts of the input sequence for each label, effectively capturing long-range dependencies between words and improving the model's ability to predict multiple labels for a single instance. To further stabilize the training process, **Layer Normalization** is applied, ensuring that the output distributions remain consistent and preventing potential issues such as vanishing or exploding gradients. Next, the **Global Average Pooling** layer aggregates the attention outputs, producing a fixed-size vector representation regardless of the input sequence length, which is particularly useful for handling variable-length customer reviews. This pooled representation is then passed through **Dense Layers**, where a fully connected layer with 128 units and ReLU activation refines the learned features, helping the model capture non-linear relationships in the data. Finally, the **Sigmoid Output Layer** produces independent probabilities for each label, utilizing the sigmoid activation function to handle the multi-label classification nature of the problem, where each instance can belong to multiple categories simultaneously. In terms of **Training Configuration**, the HiLAT was trained using the **Adam optimizer**, chosen for its efficiency in updating model weights and adaptability for complex neural networks. The model's loss function was **binary cross-entropy**, applied to handle the multi-label classification problem by optimizing the probability outputs for each label independently. A batch size of 64 was employed, allowing the model to handle sequences in manageable groups throughout training. To avoid overfitting and guarantee that the model generalizes effectively to new data, early stopping was utilized. The model underwent training for 100 epochs, with the training progress consistently observed through a 20% validation split, enabling performance assessment on unfamiliar data throughout the training stage.

Apply Models:

We gauged the model's classification quality using a combination of precision, recall, and F1-score to capture its ability to correctly identify and fully recover each label, whether frequent or rare. To complement these, we computed Hamming Loss, which measures the proportion of label assignments that differ from the ground truth across all samples. By combining these metrics, we obtain a thorough picture of how accurately and consistently the model assigns multiple labels to each customer review.

3. Theory/Calculation

3.1 Multi-label Text Classification

Multi-label text classification allocates several categories to one text, as opposed to single-label classification. Methods encompass binary relevance (transforms multi-label issues into several binary tasks), classifier chains (forecasts each label based on earlier ones), and label powerset (considers distinct label combinations as individual classes)

Numerous research works have advanced this area by employing both traditional machine learning techniques (like SVM and k-NN) and contemporary deep learning architectures such as CNNs, RNNs, and Transformers. Multi-label text classification presents particular difficulties in languages with intricate structures like Thai, owing to the absence of spaces between words and ambiguous grammatical rules. Liu et al. (2017) applied CNNs and RNNs for the multi-label classification of news articles, showcasing that deep learning models considerably enhance accuracy in multi-label tasks over conventional approaches. Nam et al. (2014) additionally highlighted that deep architectures, like neural networks, enhance the representation of intricate label dependencies. Kaur and Sharma (2020) utilized multi-label classification methods on customer reviews within the e-commerce sector, emphasizing the benefits of employing deep learning models for multi-aspect classification. Their research is significant for comprehending how multi-label approaches can be modified to assess feedback from various sources, including social media and survey answers. Devlin et al. (2019) introduced BERT, adapted for multi-label tasks, demonstrating significant

Journal of Information Systems Engineering and Management
2025, 10(56s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

improvements in text classification. Similarly, Liu et al. (2019) introduced RoBERTa, an improved version of BERT, which has shown strong generalization across various NLP tasks, including multi-label classification.

Structure of Multilabel Text Classification

Multilabel text classification is an essential method for managing data where one instance may belong to various categories at the same time. This technique applies various labels to one text, where each label indicates a distinct category or class. It is especially beneficial for tasks in which data points can be linked to multiple categories simultaneously (Zhang & Zhou, 2014).

The Multilabel classification process consists of various stages, such as preparing the data, constructing the model, training it with binary cross-entropy, establishing thresholds for predictions, and assessing the model using metrics tailored for Multilabel classification. This method is extensively utilized in practical situations, including text categorization, image labeling, and multimedia content categorization.

In multiclass classification, every instance receives one label selected from a group of exclusive categories, meaning an item can belong to only one class. In contrast, multilabel classification enables each example to possess any quantity of labels from the label set, representing cases where an instance inherently encompasses several categories. Consequently, multiclass issues yield one categorical prediction for each sample, while multilabel issues need an individual binary choice for every potential label. Assessment also varies: multiclass models are typically evaluated based on overall accuracy or single-label precision and recall, whereas multilabel models depend on metrics such as Hamming Loss and label-specific precision and recall to address multiple concurrent labels. Common multiclass tasks involve tasks like sentiment analysis focused on a single topic or identifying objects, while multilabel tasks occur in situations like assigning multiple topics to a document, diagnosing simultaneously present medical conditions, or tagging various concepts in an image (Zhang & Zhou, 2014; Tsoumakas & Katakis, 2007).

3.2 Hierarchical and Label-Wise Attention Mechanisms

Attention mechanisms have emerged as an essential element in contemporary neural architectures, enabling models to concentrate on the most vital aspects of the input sequence. In hierarchical attention networks (HAN), attention is utilized across various levels of the text framework (e.g., word and sentence levels), which proves especially beneficial for multi-label classification tasks since different labels might depend on distinct sections of the input text

Yang et al. (2016) presented the hierarchical attention network (HAN), which allocates attention weights at both the word and sentence levels, rendering it particularly suitable for tasks such as document classification and multi-label text classification. In label-wise attention, the model allocates distinct attention weights for every label, making certain that the most pertinent features of the text are highlighted for each specific label.

In multi-label tasks, hierarchical attention mechanisms focused on labels are particularly efficient as they encompass the structure of the input text and the specific importance of text features for every label. Lin et al. (2017) presented a self-attention mechanism, which has been extensively used in models such as Transformers to grasp intricate dependencies in text classification tasks.

Transformer-based architectures, especially the pretrained Transformer language models, have gained popularity for various NLP tasks. (Biswas et al., 2021) The Transformer employs self-attention mechanisms to handle input sequences. The attention mechanism enables the model to assess various sections of the input according to their significance for each label.

The HiLAT framework is designed to tackle scenarios where one text input can relate to multiple labels, such as a review that discusses both product performance and service quality. To handle these overlapping labels, HiLAT utilizes a hierarchical framework combined with attention focused on specific labels. Initially, it determines which sentences are the most pertinent for each label, and afterward, it concentrates on the most meaningful words within those chosen sentences by utilizing a distinct attention mechanism designed for each label. The equations below formalize this two-step process, as outlined by Liu et al. (2022)

Self-Attention Mechanism

In Transformer architectures, the self-attention mechanism models pairwise interactions among every token in the input sequence. The attention weights are computed according to the following formula (Vaswani et al., 2017):

Journal of Information Systems Engineering and Management
2025, 10(56s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

where:

Q, K, V denote projection vectors obtained by linearly transforming the input word embeddings.
d_k is the dimension of the K vectors.
This mechanism enables the model to prioritize the most informative tokens in the input sequence.

Label-Wise Attention Mechanism

For multilabel tasks, the model computes a distinct attention pattern for each label, allowing it to highlight different text segments depending on the label. Formally, the label-wise attention score is given by (Xiao et al., 2019):

$$\text{Attention}(Q_l, K_l, V_l) = \text{softmax}\left(\frac{Q_l K_l^T}{\sqrt{d_k}}\right)V_l$$

where l indexes a particular label, and Q_l, K_l, V_l are the query, key, and value projections tailored to that label. By computing attention weights separately for each l. the model can attend to different parts of the input sequence depending on which label it is predicting, thereby enhancing its ability to distinguish multiple labels in a single text.

Hierarchical Attention Mechanism

When processing structured texts where a document is built from multiple sentences a hierarchical approach lets the model grasp meaning at both the sentence and word layers. First, it computes attention over sentences to identify which ones carry the most weight for the overall document understanding (Yang et al., 2016):

$$h_{\text{sent},i} = \sum_{j=1}^{T_i} \alpha_{ij} h_{ij}$$

In this formulation, h_{ij} denotes the encoded vector for the j in sentence i , and α_{ij} represents the attention weight assigned to that word by the self-attention mechanism. Together, they allow the model to weigh each word's contribution when constructing a sentence-level summary.

The model then combines its sentence summaries into a single document vector by weighting each sentence according to its importance (Yang et al., 2016):

$$h_{\text{doc}} = \sum_{i=1}^{N} \beta_i h_{\text{sent},i}$$

Here, h_{doc} represents the overall document embedding, and β_i s the attention coefficient indicating how much weight sentence i contributes to that final representation.

Multilabel Classification with Sigmoid Output

After computing attention, the model uses the sigmoid activation function to make multilabel predictions. The prediction for each label is given by:

$$\hat{y}_l = \sigma(W_l h_l + b_l)$$

where:

$\hat{y}_l$is the predicted value for label l.
h_l is the output from the attention mechanism for label l.
W_l and b_l are the learnable parameters for label l.
σ denotes the logistic (sigmoid) function, which squashes its input into the interval [0, 1].

Journal of Information Systems Engineering and Management
2025, 10(56s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Algorithm 1: HiLAT for Multi-Label Text Classification

1: Initialize model parameters with hierarchical structure, word-level, and sentence-level attention layers.
2: Initialize label-wise attention mechanisms for each label.
3: Initialize attention vectors Q, K, V with random weights for word-level attention.
4: Initialize exploration rat ϵ (if applicable for dynamic learning or exploration strategies).

5: **for** document $d = 1$ to D (*Loop over each document*)
6: Preprocess document d into sentences s_i and words w_{ij} within each sentence.
7: **for** $t = 1$ to T (*Loop over each word in sentence*)
8: Apply word-level attention: $h_{\text{sent},i} = \sum_{j=1}^{T_i} \alpha_{ij} h_{ij}$ where α_{ij} is attention weight for word w_{ij}
9: Store word-level representation $h_{\text{sent},i}$ for sentence s_i.
10: **end for**

11: Compute sentence-level attention: $h_{\text{doc}} = \sum_{i=1}^{N} \beta_i h_{\text{sent},i}$ where β_i is the attention weight for sentence s_i.
12: **for each label** l (*Apply label-wise attention*)
13: Compute label-specific attention:

$$\text{Attention}(Q_l, K_l, V_l) = \text{softmax}\left(\frac{Q_l K_l^T}{\sqrt{d_k}}\right) V_l$$

14: Generate prediction for label l: $\hat{y}_l = \sigma(W_l h_l + b_l)$

15: **end for**

16: Update weights by minimizing the loss $L(\theta)$ based on predictions and actual labels for document d.

17: Every C steps, update the model parameters for hierarchical and label-wise attention mechanisms based on loss minimization.
18: Update exploration rate ϵ if adaptive exploration is used.
19: **end for**

20: **return** Final predictions $\hat{y}_l$ for all labels across all documents.

In algorithm 1, symbols and components are defined as follows:

D is the set of documents, where each document consists of multiple sentences and words structured hierarchically. The attention vectors Q, K, V are used in the self-attention mechanism, initialized with random weights to capture relationships between words within sentences. For each sentence i the model computes a sentence-level representation $h_{\text{sent},i}$ by applying word-level attention to each word in the sentence, where h_{ij} is the representation of word j in sentence i.

The attention weight α_{ij} is assigned to each word j in sentence i, calculated through the self-attention mechanism to focus on the most relevant words. These word representations are then aggregated to form the sentence representation $h_{\text{sent},i}$. Subsequently, the model computes the document-level embedding h_{doc}, which aggregates sentence-level representations across all sentences in the document. The attention weight β_i is the attention weight assigned to sentence i, highlighting important sentences for document-level understanding.

In the multi-label classification task, each label l has a specific focus within the text. To capture this, label-specific attention vectors Q_l, K_l, V_l are introduced, allowing the model to focus on different parts of the text for each label. The

Journal of Information Systems Engineering and Management
2025, 10(56s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

predicted probability for each label l, represented as $\hat{y}_l$ is calculated by passing the label-specific attention output through a sigmoid activation function, which maps the result to a probability range.

To improve prediction accuracy, the model includes learnable parameters W_l and b_l for each label l, which transform the label-specific attention output into a final prediction. The model's performance is optimized by minimizing a loss function $L(\theta)$, which measures the difference between the predicted probabilities and the actual labels. This loss is minimized to update model weights and enhance prediction accuracy.

During training, parameter updates occur every C steps, ensuring that model parameters are refreshed periodically to improve learning. If adaptive exploration is applied, an exploration rate ϵ s used to balance exploration and exploitation during training, helping the model dynamically adapt its learning strategy.

Algorithm 1 sets up hierarchical attention layers and label-specific attention mechanisms for handling multi-label classification. It analyzes each document by utilizing word-level attention to identify key words in sentences, followed by sentence-level attention to condense the significance of sentences for the whole document. For every label, a particular attention mechanism concentrates on pertinent sections of the document, producing predictions specific to each label. The model periodically adjusts its parameters to reduce the loss, with the goal of enhancing classification accuracy for all labels. The end result is a collection of forecasts for each label across all documents, with each forecast signifying the likelihood of that label's presence.

3.2 Handling Imbalanced Datasets in Multi-Label Classification

Imbalanced datasets present a major difficulty, particularly in multi-label classification, since certain labels are inadequately represented. If not dealt with, this may result in inadequate performance on minority labels. Methods to tackle imbalance consist of resampling (increasing samples for minority classes or decreasing samples for majority classes), cost-sensitive learning (imposing greater misclassification penalties on minority classes), and data augmentation (creating synthetic data for underrepresented categories).

For Thai text categorization, where certain labels may have limited representation, imbalance strategies are essential. Charte et al. (2015) examined approaches to tackle the imbalance in multi-label learning, including methods such as SMOTE, which has been tailored for multi-label tasks to address class imbalance. In a similar vein, Chawla et al. (2002) proposed the SMOTE algorithm, which has established itself as a key technique for tackling imbalance in classification problems.

Performance Metrics

In assessing multilabel text classification models, it is crucial to utilize performance metrics that are capable of addressing the unique characteristics of multilabel classification, where each instance may possess several labels at once. Conventional single-label metrics such as accuracy and precision must be modified or expanded for the multilabel context. Presented here are several important performance metrics frequently utilized in multilabel text classification (Zhang & Zhou, 2014).

Precision:

$$precision_j = \frac{TP_j}{TP_j + FP_j}$$

where:

TP_j is the number of true positives for label j
FP_j is the number of false positives for label j
The macro-averaged precision, which is the average precision across all labels, is given by:

$$precision_{macro} = \frac{1}{L}\sum_{j=1}^{L} precision_j$$

where L represents the total number of labels.

Journal of Information Systems Engineering and Management
2025, 10(56s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Recall:

For a single label j, recall is defined as:

$$Recall_l = \frac{TP_j}{(TP_j + FN_j)}$$

where: FN_j is the number of false negatives for label j
The macro-averaged recall is given by:

$$Recall_{macro} = \frac{1}{L}\sum_{j=1}^{L} Recall_j$$

F1 Score:

For a single label j, the F1 score is defined as:

$$F1_j = \frac{2 \times precision_j \times recall_j}{recall_j + precision_j}$$

The macro-averaged F1 score is given by:

$$F1_{macro} = \frac{1}{L}\sum_{j=1}^{L} F1_j$$

Hamming Loss:

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

where N is the number of samples, L is the number of labels, and 1 is the indicator function.

RESULTS

This study demonstrates the development and evaluation of a HiLAT model for multi-label classification of customer feedback from Thai banks. The model was designed to address the challenges posed by imbalanced datasets, where certain categories are underrepresented. By utilizing hierarchical attention mechanisms, the model emphasizes key parts of customer feedback tied to each label, offering an in-depth understanding of customer viewpoints.

Table 1 presents the performance metrics of the HiLAT model, utilized to categorize Thai banking customer reviews into eight distinct groups. The table displays essential performance metrics such as precision, recall, F1-score, and Hamming Loss to assess the model's effectiveness in recognizing pertinent customer feedback across various categories.

Category	Precision	Recall	F1-Score	Hamming Loss
Image	0.470	0.512	0.490	
Product and Services	0.609	**0.917**	**0.732**	
Facility and Supporter	0.367	0.476	0.415	
Staff	0.635	0.620	0.627	

Journal of Information Systems Engineering and Management
2025, 10(56s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Category	Precision	Recall	F1-Score	Hamming Loss
Timing	0.563	0.639	0.599	
Accessibility	0.804	0.661	0.725	
Chatbot	0.366	0.431	0.396	
Other	0.553	0.541	0.547	
Macro avg	0.572	0.639	0.597	
HiLAT (Overall)				**0.233**

The HiLAT model excels in identifying labels with numerous training examples, notably in Product and Services as well as Accessibility. For Products and Services, it achieves a recall of 0.917 and an F1-score of 0.732, showing it rarely misses service-related comments, although the precision (0.609) indicates the presence of some false positives. Accessibility annotations are accurately documented, demonstrating a precision of 0.804, a recall of 0.661, and an F1 score of 0.725, which suggests that the hierarchical attention effectively targets the most relevant text portions for these frequent labels. Conversely, categories that lack sufficient representation, such as Chatbot, Facility and Supporter, and Image, still present challenges. Feedback from the chatbot shows a low F1 score of 0.396 (precision 0.366, recall 0.431), highlighting the occurrence of missed instances and false positives. Facility and Supporter achieve F1 0.415 (precision 0.367, recall 0.476), indicating that the model often detects real cases rather than overlooking false positives. Observations about images perform somewhat better but still modestly, with an F1 score of 0.490 (precision 0.470, recall 0.512), highlighting the difficulties of detecting visual-interface issues in unstructured text. For mid-frequency labels like Staff and Timing, HiLAT finds a balance: Staff yields an F1 score of 0.627 (precision 0.635, recall 0.620), while Timing achieves an F1 score of 0.599 (precision 0.563, recall 0.639). The Other category, which includes diverse feedback, achieves an F1 score of 0.547, reflecting its diverse nature. Overall, the macro-averaged F1 score of 0.597 and a Hamming Loss of 0.233 indicate that HiLAT effectively recognizes the majority of relevant labels—especially in well-represented classes—while still misclassifying nearly one in four instances. Targeted techniques such as data augmentation or oversampling for underrepresented groups could further diminish this performance gap.

DISCUSSION

This research utilized the HiLAT model to categorize Thai banking customer feedback into eight groups amid significant class imbalance and distinct linguistic intricacy. The assessment findings (Table 4.4) indicate that HiLAT is very effective for well-represented labels: Product and Services obtained an F1-score of 0.732 (precision 0.609, recall 0.917), while Accessibility achieved an F1-score of 0.725 (precision 0.804, recall 0.661). These robust findings suggest that the hierarchical attention mechanism effectively identifies and highlights the most pertinent sentences and tokens for frequent labels, consistent with previous studies on attention-based models (Vaswani et al., 2017). Mid-frequency categories like Staff (F1 0.627) and Timing (F1 0.599) were managed fairly effectively, demonstrating a balanced precision-recall trade-off (Staff: 0.635/0.620; Timing: 0.563/0.639). The Other category, which includes varied feedback, recorded an F1 of 0.547—less than the common classes but still reflective of HiLAT's capability to generalize across diverse inputs.

Achievement in underrepresented categories continues to be a hurdle. The feedback from the chatbot resulted in an F1 score of 0.396 (precision 0.366, recall 0.431), while Facility and Supporter achieved an F1 score of 0.415 (precision 0.367, recall 0.476). Even with the use of class-weighted loss, these minority labels still faced a lack of adequate training samples. Image feedback also turned out to be challenging (F1 0.490), indicating that remarks regarding visual or interface elements need extra input or customized enhancements. In total, HiLAT reached a macro-averaged F1-score of 0.597 and a Hamming Loss of 0.233, signifying that the model accurately predicts most labels while misclassifying roughly 23% of label assignments on average. These findings indicate that hierarchical, label-wise attention paired with class weighting significantly enhances performance compared to unweighted or non-hierarchical baselines, while also underscoring the ongoing deficiency in minority class recall. To tackle these

shortcomings, upcoming efforts should investigate sophisticated data-level techniques—like oversampling (for instance, multi-label SMOTE), back-translation, or focused paraphrasing—to enhance minority class instances. Techniques at the model level, such as focal loss or supplementary adversarial training, could enhance the recall of minority classes. Moreover, adding domain-specific embeddings developed on extensive Thai financial texts might assist the model in detecting finer linguistic signals associated with less represented categories.

CONCLUSION

The Hierarchical Label-Wise Attention Transformer (HiLAT) demonstrates significant potential for multi-label classification of Thai banking customer feedback, attaining solid results in well-represented categories (F1 0.732 for Products & Services; F1 0.725 for Accessibility) and an overall macro-average F1 of 0.597 across all eight labels. The model's tiered attention successfully captures context specific to labels, while the class-weighted loss reduces—but does not completely remove—the negative consequences of significant class imbalance. A Hamming Loss of 0.233 verifies the system's overall dependability, with accurate label assignments in almost 77% of instances. Nonetheless, results in minority categories (Chatbot F1 0.396; Facility & Supporter F1 0.415; Image F1 0.490) suggest that additional improvements are required. Future studies ought to concentrate on enhancing limited labeled data via augmentation and oversampling, testing different loss functions, and utilizing domain-specific embeddings to boost recall for underrepresented categories. These actions will be crucial for enhancing HiLAT's overall utility in practical applications—like banking, healthcare, and e-commerce—where thorough, balanced feedback analysis is vital for informed decision-making and service enhancement.

ACKNOWLEDGMENTS

We thank Chiang Mai University for their support. Special thanks to our linguist team for manual annotation of the dataset. We also acknowledge the use of AI tools to enhance writing clarity and perform grammar checking.

REFRENCES

[1]Biswas, B., Pham, T.-H., & Zhang, P. (2021). TransICD: Transformer-based code-wise attention model for explainable ICD coding. In A. Tucker, P. Henriques Abreu, J. Cardoso, P. Pereira Rodrigues, & D. Riaño (Eds.), *Artificial Intelligence in Medicine: AIME 2021* (Lecture Notes in Computer Science, Vol. 12721, pp. 471–481). Springer. https://doi.org/10.1007/978-3-030-77211-6_56

[2]Buda, M., Maki, A., & Mazurowski, M. A. (2018). A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks, 106*, 249–259. https://doi.org/10.1016/j.neunet.2018.07.011

[3]Chantree, S., Singpeng, S., & Chaikul, P. (2021). Thai word segmentation using deep learning: Challenges and approaches. *International Journal of Computer Applications, 174*(21), 1–6. https://doi.org/10.5120/ijca2021921775

[4]Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16*, 321–357. https://doi.org/10.1613/jair.953

[5]Charte, F., Rivera, A. J., del Jesus, M. J., & Herrera, F. (2015). Addressing imbalance in multi-label classification: Measures and random resampling algorithms. *Neurocomputing, 163*, 3–16. https://doi.org/10.1016/j.neucom.2014.08.091

[6]Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 4171–4186). https://doi.org/10.18653/v1/N19-1423

[7]He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering, 21*(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

[8]Kaur, H., & Sharma, S. K. (2020). Multilabel text classification for analyzing customer feedback in e-commerce industry using deep learning techniques. *Journal of Web Engineering, 19*(1), 1–16.

[9]Lin, Z., Feng, M., dos Santos, C. N., Yu, M., Xiang, B., Zhou, B., & Bengio, Y. (2017). A structured self-attentive sentence embedding. In *Proceedings of the 5th International Conference on Learning Representations (ICLR)*. https://arxiv.org/abs/1703.03130

[10]Liu, L., Zhang, Y., Wang, M., Wang, Y., Sun, J., & Wu, Y. (2022). Hierarchical label-wise attention transformer model for explainable ICD coding. *Journal of Biomedical Informatics, 133*, 104161. https://doi.org/10.1016/j.jbi.2022.104161

[11]Liu, P., Qiu, X., & Huang, X. (2017). Recurrent neural network for text classification with multi-task learning. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI)* (pp. 2285–2291). https://doi.org/10.24963/ijcai.2017/318

[12]Liu, Z., Lin, W., Shi, Y., & Zhao, J. (2021). A robustly optimized BERT pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics* (pp. 1218–1227). Hohhot, China: Chinese Information Processing Society of China.

[13]Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv Preprint*, arXiv:1907.11692. https://arxiv.org/abs/1907.11692

[14]Nam, J., Kim, J., Mencía, E. L., Gurevych, I., & Fürnkranz, J. (2014). Large-scale multi-label text classification—Revisiting neural networks. *Machine Learning, 93*(1), 41–75. https://doi.org/10.1007/s10994-013-5333-3

[15]Tsoumakas, G., & Katakis, I. (2007). Multi-label classification: An overview. *International Journal of Data Warehousing and Mining, 3*(3), 1–13. https://doi.org/10.4018/jdwm.2007070101

[16]Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems, 30*, 5998–6008.

[17]Xiao, X., Zhang, S., Wang, L., & Li, J. (2019). A label-wise attention mechanism for multilabel text classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 4567–4578). https://doi.org/10.18653/v1/D19-1463

[18]Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., & Hovy, E. (2016). Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 1480–1489). https://doi.org/10.18653/v1/N16-1174

[19]Zhang, M. L., & Zhou, Z. H. (2014). A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering, 26*(8), 1819–1837. https://doi.org/10.1109/TKDE.2013.39