International Journal of Computational and Experimental Science and ENgineering (IJCESEN)

Vol. 11-No.3 (2025) pp. 4555-4564
http://www.ijcesen.com

ISSN: 2149-9144

Research Article

Thai Text-based Classification for Small and Imbalanced Dataset

Chartwut Thanajiranthorn[1], Warawut Chosungnoen[2], Thippawan Saenkham[3], Nuttapol Saenkham[4*]

[1]Faculty of Science, Buriram Rajabhat University, Thailand
Email: chart@gmail.com - **ORCID:** 0000-0002-5247-7150

[2]Faculty of Science, Buriram Rajabhat University, Thailand
Email: warawu2t@gmail.com - **ORCID:** 0000-0002-5247-7250

[3]Faculty of Science, Buriram Rajabhat University, Thailand
Email: thip@gmail.com - **ORCID:** 0000-0002-5247-7350

[4]Faculty of Science, Buriram Rajabhat University, Thailand
* **Corresponding Author Email:** nuttapol.sk@bru.ac.th - **ORCID:** 0000-0002-5247-7450

Article Info:

DOI: 10.22399/ijcesen.3015
Received : 05 May 2025
Accepted : 19 June 2025

Keywords

Text Classification
Imbalanced Data
Data Augmentation
Thai Text

Abstract:

Insufficient and imbalanced data are critical issues for developing a text-based classification because they hinder its performance, including underfitting and inaccuracy. This work presents a method to apply text data generation methods to systematically increase data quantity and leverage the number difference between classes. Data augmentation methods such as synonym replacement and data synthesis methods are exploited to generate additional Thai text data based on existing data. For training towards classification models, generated data and original data are merged to solve small and uneven dataset issues and improve classification performance. From experimental results, the overall F1 score obtained from all datasets with generated data is significantly higher than the baseline if the original dataset is used. The improvement of the major categories is to gain higher precision, while the minor categories have had their recall greatly improved. For a single-generation method, data augmentation by synonym replacement produces the highest F1 score of 0.83. The combination of synonym replacement and GPT4 yields the best result in average classification performance for a 0.86 F1 score.

1. Introduction

Text mining is the process of analyzing information and patterns from unstructured textual data to capture key concepts and hidden relationships using natural language processing (NLP) techniques. It is widely used in several applications, including spam detection, sentiment analysis, and document categorization. This method has been applied in various contexts, including the use of advanced algorithms for improved performance [1]. One of the most frequently used tasks for text mining is text classification. It focuses on categorizing text documents or sentences into predefined categories based on given content. Text classification is widely used in several applications, including spam detection, sentiment analysis, topic labeling, and document categorization. Text classification automates the process of sorting and labeling text, which otherwise requires a tedious manual endeavor. It can greatly benefit management and policymaking by providing insights, gauging trends, and enabling proactive measures.Despite many benefits, text classification requires an appropriate amount of data and some degree of text quality to be effective and accurate. Insufficient data in text classification, especially multiclass classification, leads to failure to learn the relevant features representing the classes as underfitting [2]. On the other hand, the model based on insufficient data might be overfitting the training dataset by learning noise instead of the underlying features, resulting in poor performance with unseen data [3]. Furthermore, insufficient training data often comes with imbalanced data, as one or more classes have significantly fewer instances than the other classes. For the imbalanced dataset, some classes are underrepresented; thus, the classification model becomes biased towards the more frequent classes and leads to inaccurate categorization [4,5]. Hence, applying text classification to the low amount of data

and the imbalanced data needs extra solutions to accomplish the task effectively.In Thailand, research funds and their allocated amount are based on the developed national strategy which is employed as the country's goal for sustainable national development in accordance with the principle of good governance. The scale of funding institutes can be at the national, regional, provincial level, or university-based research project level (funding from university/college). The focus funding aspects are thus different according to the scale. For the university-based research project scale, the number of research proposals can be between 40 to 60 topics annually. To classify the research proposals into categories of the national strategy, proposal owners need to select one of the strategies manually. The selected categories however are occasionally incorrect, and one of the complaints from users is the difficulty to appropriately select one as there are many categories, and their details are lengthy and hard to incomprehensible for other expertise fielders.This work aims to develop a text classification for the Thai national strategy based on the given executive summary in Thai. The core challenges are the low number of training datasets and imbalanced training data. To solve the issues, data augmentation and data synthesis are applied to increase the amount of training data. For classification, several machine learning techniques are used in an ensemble fashion to achieve the highest performance.

2. Background

There are few research articles focusing on using text mining techniques for detecting inappropriate Thai text on social networking platforms. In 2018, Arreerard and Senivongse (2018) presented a method for detecting and identifying defamatory content from Thai texts on social media [6]. They applied several machine learning techniques for comparison, such as Support Vector Machines (SVM) and Naive Bayes. The dataset is a collection of about 1,000 posts made on social media, and 446 defamatory statements among the posts are annotated. Their experiment results show that their best performance is from SVM with features of word n-grams and character n-grams for F1 score of 0.64. Hemtanon et al. proposed using text classification for online harassment detection in social media messages [7]. They annotated 12,000 textual posts for binary classification of malicious intent to insult and threaten other users. The data were trained using the Decision Tree method for classification and extracted keywords that represented harassment expressions. Based on their findings, the classification method demonstrated an average performance with an F1 score of 0.78.Another common task for Thai text classification is sentimental analysis. The sentiment analysis classification is to identify the feelings or intention of the text poster into a category of positive, negative, and neutral feelings. Chumwatana (2015) applied text mining to Thai online customer reviews SMOTE available on social media and websites for sentiment analysis and classification [8]. The proposed method was based on the integration of Thai word extraction and identifying the word as positive or negative using classification. In 2020, Tanantong et al. collected 4,608 textual posts from several social media platforms, including Facebook, Instagram, and Pantip to classify opinions on the mobile network operator [9]. The texts were assigned negative and positive labels and trained for a classification model. Their highest accuracy was an 87.7$\%$ accuracy score. Khamphakdee and Seresangtakul presented their work on sentiment analysis methods using advanced deep learning techniques such as CNN, LSTM, and GRU [10]. The dataset was 25,398 customer text reviews collected from hotel booking platforms. The Word2Vec model was applied to build word embedding dimensions, while Delta TF-IDF was exploited to extract features from text data to use in the deep learning process. FastText and BERT pre-trained models were then used to perform the sentiment classification.Lastly, there are research studies aimed at identifying Thai texts into specific topics. This topic classification can be a news category, an academic domain category, or any category as needed. Klaithin and Haruechaiyasak applied machine-learning techniques to identify texts about traffic on Twitter [11]. They collected 24,779 text messages and preprocessed them by removing duplicate tweets and deleting unrelated content. The texts then were labeled with topics such as road accidents, traffic news, help requests, and feelings about traffic. In 2018, Wongsap et al. presented their work on the classification of clickbait headline news [12]. Their dataset consisted of 5,000 headline texts with positive and negative labels of clickbait. Frequently used classifiers such as Decision Tree, Support Vector Machine, and Naive Bayes were applied to develop a classification model with term frequency and word sequence (n-gram) as features.From the reviews, Thai text classification models can be developed using traditional machine learning techniques and advanced deep learning techniques. The criteria are the amount of training data and the characteristics of the text. As advanced deep learning techniques require a larger training dataset, some datasets are not applicable to such techniques. For

classifying features, term frequency and word sequence are often used to represent textual information and show good performance. So far, the results of Thai text classification are acceptable, especially binary classification. However, the existing studies usually had more than a thousand instances for the training set, which is sufficient for classifiers to learn their distinguishable features. None of the studies attempt multi-class classification using datasets with the average instances per class under 100 training instances. Furthermore, the issue of imbalanced Thai text data has not been discussed much since their applied datasets were fairly distributed and sufficiently compact to represent the classes. Thus, this work aims to solve the insufficient and imbalanced issue of Thai text classification so we can utilize classification on small and imbalanced text data.

3. Development of Classification Models

This work aims to categorize an executive summary of a research project proposal into Thai research strategy categories. The summary of the proposal is in Thai and processed using basic NLP methods. The processed text data with the category label are trained into a classification model using multiple supervised learning techniques such as Decision Tree, Gaussian NB, and Logistic Regression. Then, a voting scheme is exploited to make use of the decisions made by all models. An overview of the system is illustrated in Figure 1.

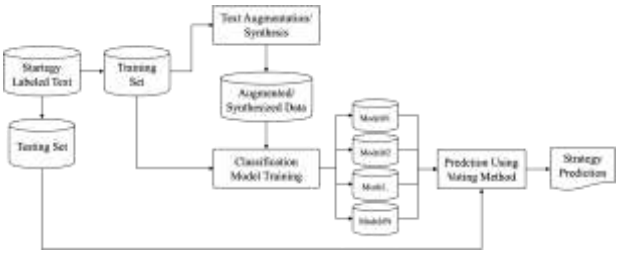

Figure 1. *Overview of the processes*

From Figure 1, there are four main parts in the process of overview data collection to acquire strategy-labeled data for training and testing datasets, text augmentation and synthesis to solve insufficient and imbalanced data issues, classification model training based on selected techniques and voting methods to collectively use all generated models towards prediction of a strategy category.

3.1. Data Collection

Data in this work are a collection of an executive summary in Thai as unstructured text data from research project proposals within Buriram Rajabhat University between the fiscal years 2022-2024 as a university-based research project scale. The data were labeled with the strategy category by project owners and were approved by the funder committee. The total number of the collected data is 251 documents. The number of national strategy categories is 6, as follows:

- National Security (NS)
- National Competitiveness Enhancement (CE)
- Human Capital Development and Strengthening (DS)
- Social Cohesion and Just Society (CJ)
- Eco-Friendly Development and Growth (ED)
- Public Sector Rebalancing and Development (RD)

The nation's security, prosperity, and sustainability are the focal points of the six distinct categories. Data statistics of the original dataset by category label are given in Figure 2.

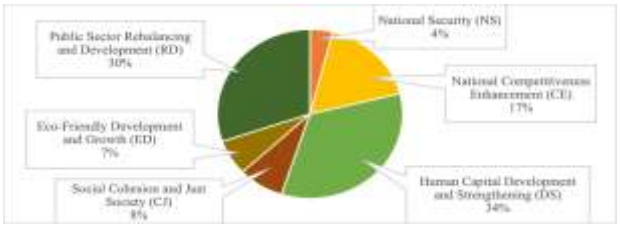

Figure 2. *Data statistics of the collected text data used in this research*

From the information, this labeled corpus is imbalanced as some categories, such as DS and RD have more documents than the rest, especially NS which has only 10 documents to represent the category. The maximum and minimum number of instances are 85 and 10, respectively. Moreover, the total number of documents is 251, with an average of 41.83 instances per category, which is considerably low for the task of text classification.

3.2. Data Augmentation and Synthesis

In this work, the original dataset incurs two main issues: insufficiency and imbalance. Thus, we apply two methods, including data augmentation and data synthesis, to solve the issues. Data augmentation is a method to increase the diversity of a dataset by creating modified versions of existing data while keeping the core of the data [13]. For classification, the method helps to improve the model's ability to

generalize by exposing it to a broader variety of data, especially for classes with much fewer instances than other classes. Differently, data synthesis is to generate entirely new text data that does not exist in the original dataset based on the characteristics of the original data. Both methods are useful to solve the insufficiency and imbalance issues, but the created data is based on different concepts and may affect the outcome of the classification task. Thus, we plan to conduct a comparative experiment between the two and use the most effective method for the classification task. For data augmentation, there are many techniques, as follows:

- synonym replacement: replacing words in the text with their synonyms [14].
- random swap: swapping the positions of words in the text randomly [15].
- back translation: using machine translation to translate the text into another language and translate the translation back into the original language [16].
- sentence shuffling: for longer texts or paragraphs only, shuffling the sentences into different order [17].

These techniques, however, require some linguistic information or resources to accomplish. For example, synonym replacement needs a wordlist from a synonym dictionary to replace the words, while back translation and sentence shuffling require a marking of sentence boundary (such as the use of a full-stop symbol in English), and both sentence and word boundary for random swap to perform correctly. For this work, as the Thai language does not have an explicit sentence marker, and the opened tool for Thai sentence segmentation is not generally accurate, we hence cannot afford the back translation and sentence shuffling technique for data augmentation. Thus, we solely apply the synonym replacement technique for augmenting text data for this work. The synonym list in this work is based on the electronic version of the Royin Thai Dictionary (2011 edition) [18] which is the most referable official Thai monolingual dictionary. For the replacement scheme, in the context of the research proposal, we focus on replacing specific parts of speech, including verbs and adjectives, in the text as candidates for replacement since nouns in the text may hold technical meaning and often are a part of compound nouns or coined names. For the number of words to be replaced, we set it to 20% of all candidate words in a document. The candidate words are then randomly replaced with one of the synonym terms from the dictionary.For data synthesis, there are two famous methods including SegGan, SMOTE for text, and Generative Pre-trained Transformer 4 (GPT-4). SegGan is an enhanced method of GAN framework designed for generating sequences for text generation. The core technique is to generate sequences of text by sampling from a probability distribution and distinguish between real and synthetic data using advanced learning approaches such as RNNs, and LSTMs. Then, SegGan applies reinforcement learning to train the generator while considering feedback from evaluation of distinguishing in the form of rewards to guide the generator to produce more realistic sequences. This method is capable of generating realistic text data but requires high computation and high amount of training data for tuning. Thus, it is not suitable for our context. SMOTE for text is to convert words in text into a numerical representation in a vector (Word2Vec) and find its k-nearest neighbors in the embedding space to generate synthetic samples by interpolating between the embeddings of the sample and its neighbors. For example, for each synthetic sample, we randomly selected a neighbor from the k-nearest neighbors. The new word vector was calculated as follows: new_vector = original_vector + alpha * (neighbor_vector - original_vector). where alpha is a random number between 0 and 1.For GPT-4, we utilized the OpenAI API with the `gpt-4-0613` model. The following prompt structure was used to generate synthetic research proposal summaries for the [National Security] category: “Generate a concise summary (approximately 150-200 words) of a research proposal that aligns with the Thai national strategy category of [National Security]. The summary should address the following: 1) the research problem, 2) the proposed methodology, 3) the expected outcomes, and 4) the potential impact on national security. The tone should be formal and academic. Generate a new, unique summary.”. The API parameters were set as follows: temperature=0.7, top_p=0.95, frequency_ penalty=

0.2, presence_penalty=0.2, max_tokens=256.

In this work, we thus choose to apply 2 techniques SMOTE for text and GPT-4 from OpenAI to generate synthetic text data to increase the amount of the training data.

3.3. Classifier Models

Since the training data in this work is small, traditional techniques are chosen to develop models for classification. In this work, we choose 5 frequently used supervised learning techniques as follows.

- **ID3-Decision tree**

○Description: a flowchart-like tree structure where internal nodes represent decisions based on features, branches represent the outcome of these decisions, and leaf nodes represent the classes for prediction. The main concept of the method is to split the data into different sets based on the value of input features.
○Parameter setting: ID3 (Iterative Dichotomiser 3) algorithm is chosen. Maximum depth is set to unlimited. The minimum sample split is set to 2. The function to measure the quality of a split is based on information gain using entropy.

- **Gaussian Naive Bayes**

○Description: a probabilistic classifier based on Bayes' theorem by assuming the likelihood of the word features is normally distributed. The text data are calculated and represented using a vector of TF-IDF scores for words. The classifier learns the probability distribution of words given each class and makes a prediction of new text data based on the learned probabilities.
○Parameter setting: Multinomial Naive Bayes (MNB) approach, TF-IDF calculation to represent word features, and training a classifier on the vectorized text data to learn the probability distribution of words given to each class.

- **Logistic regression**

○Description: A linear model used for binary classification tasks that models the probability of a binary outcome using a logistic function. Logistic regression models the probability that a document instance belongs to a class using the sigmoid function. Each document is represented in the form of a vector of TF-IDF scores of words.
○Parameter setting: an instance of an executive summary is vectorized with TF-IDF scores of words in the summary. The approach for classification is One-vs-Rest (OvR) as each classifier is trained to distinguish between one class and all other classes. A threshold for classification is set to 0.5.

- **Support Vector Machine (SVM)**

○Description: a supervised machine learning algorithm aims to find the optimal hyperplane that maximally separates the data points of different classes in a high-dimensional space.
○Parameter setting: non-linear SVM with sigmoid kernel is chosen. C value is set to '1.0'. gamma is set to scale.

- **Neural network**

○Description: a computational model inspired by the way biological neural networks in the human brain process information. It consists of interconnected layers of nodes, where each node processes input data and passes the result to the next layer.
○Parameter setting: vocab_size = 5000, max_length=100, embedding_dim = 50, batch_size = 16, epochs = 10, learning rate = 0.001)

Each technique has a different focus on classifying. Thus, it is possible that the classification result may vary based on the characteristics of the method. With all classification models, combining the predictions of multiple models is used to produce a final prediction. Since each model may classify documents into different categories, we apply the voting mechanics of ensemble learning to find the most classified category [19].

4. Experiments and Results

In this section, we evaluate the performance of the proposed method in two aspects. First, we examine the performance of the solutions to insufficient and imbalanced data by using the collected dataset. Second, the new and unknown documents were tested for classification, and the classification results were assessed by the project owner to determine whether they were correct or not.

4.1. Performance Evaluation of Data Generation

This experiment is to evaluate the classification using different techniques for solving insufficient and imbalanced data. The dataset statistic is as shown in Figure 2. The total number of documents is 251, with an average of 41.83 instances per category, while the minimum and maximum number of instances are 10 and 85, respectively. There are 3 techniques of data generation to solve data issues for comparison, including synonym replacement (SR) for data augmentation, SMOTE for Text (SfT), and GPT-4 from OpenAI (GPT) for data synthesis. For classification, all classifying techniques mentioned in 3.3 are applied with majority voting for ensemble learning. 5-fold cross validation is used to split training and testing data for model evaluation. The metrics used for performance evaluation in this work are Precision (P), Recall (R), and F1 score were calculated using the following equations:

$$P = \frac{TruePositives}{TruePositives + FlasePositives}$$

$$R = \frac{TruePositives}{TruePositives + FlaseNegatives}$$

$$F1 = 2\frac{P \cdot R}{P + R}$$

The true positives are the number of documents that are correctly classified. The false positives are the number of documents that are incorrectly predicted as a targeted category but actually belong to another category. The false negatives are the number of documents that are incorrectly predicted as one of the other categories but actually belong to the targeted category.For setting, the baseline is the original data. In terms of data generation in model training process, the instances of each category are to be generated to 100 instances to prevent the imbalanced issue and to increase the data number for sufficient training. The setting of data generation for this experiment is as follows.

- SR: using only synonym replacement for generation.
- SfT: using only SMOTE for Text for generation.
- SfT: using only GPT-4 from OpenAI for generation.
- SR-SfT: using synonym replacement and SMOTE for Text for generation of a 1:1 ratio of data.
- SR-GPT: using synonym replacement and GPT-4 from OpenAI for Text for generation of a 1:1 ratio of data.
- SfT-GPT: using SMOTE for Text and GPT-4 from OpenAI for generation of a 1:1 ratio of data.
- ALL: using SR, SfT, and SfT for generation equally.

For example, the category with 20 instances for training data will get the addition of 40 augmented data from SR and another 40 synthetic data from SMOTE for Text in the SR+SfT method, while the category with 50 instances will obtain 25 generated data from each. In the case of the remainder from the division, the remainder priorities the SR method and the GPT method, respectively. Evaluation results of precision, recall, and F1 score by comparing categories and techniques are given in Table 1. Table 2 gives the overall performance of data generation techniques

***Table 1**. Evaluation results based on categories and data generation techniques.*

	NS			**CE**			**DS**			**CJ**			**ED**			**RD**		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
Baseline	1.00	0.10	0.18	0.76	0.85	0.81	0.60	1.00	0.75	0.75	0.16	0.26	0.80	0.25	0.38	0.56	0.98	0.71
SR	0.83	0.56	0.67	0.79	0.89	0.84	0.81	0.93	0.87	0.69	0.73	0.71	0.71	0.77	0.74	0.76	0.93	0.84
SfT	0.67	0.50	0.57	0.82	0.86	0.84	0.77	0.91	0.84	0.67	0.67	0.67	0.67	0.62	0.64	0.75	0.93	0.83
GPT	0.71	0.63	0.67	0.84	0.84	0.84	0.78	0.91	0.84	0.64	0.60	0.62	0.75	0.64	0.69	0.74	0.95	0.83
SR-SfT	0.86	0.67	0.75	0.80	0.91	**0.85**	0.82	0.96	0.88	0.80	0.71	**0.75**	0.77	0.71	0.74	0.78	0.97	0.87
SR-GPT	0.86	0.67	0.75	0.82	0.86	0.84	0.82	0.97	**0.89**	0.86	0.67	**0.75**	0.85	0.73	**0.79**	0.81	0.98	**0.89**
Sft-GPT	1.00	0.60	0.75	0.89	0.79	0.84	0.80	0.93	0.86	0.75	0.75	**0.75**	0.71	0.77	0.74	0.76	0.97	0.85
ALL	0.88	0.78	**0.82**	0.86	0.84	**0.85**	0.80	0.96	0.87	0.80	0.71	**0.75**	0.85	0.73	**0.79**	0.80	0.98	0.88

Table 1 presents the detailed evaluation results for each national strategy category using different data generation techniques. The table shows the Precision, Recall, and F1-score for each category and each technique, allowing for a granular comparison of performance. Precision measures the proportion of correctly classified instances among those predicted to belong to a given category. Recall measures the proportion of actual instances of a category that are correctly identified. The F1-score is the harmonic mean of Precision and Recall, providing a balanced measure of overall performance.The results from Table 1 indicate that imbalanced data play a crucial role in classification performance, as shown in the evaluation results of the baseline, as categories with a larger number of documents, including CE, DS, and RD, obtain a high recall and F1 score over 0.7, while categories with a low number of documents, such as NS, CJ, and ED, suffer from a low recall and F1 score. These show that the model is biased towards the categories with more data and does not see enough examples of the minor categories to learn their distinguishing features. With data generation techniques,

classification performance is noticeably improved for all applied metrics. F1 scores from minor categories are all increased by over 0.40, especially for the NS category, where their recall and F1 scores improve from 0.10 and 0.18 to 0.56 and 0.67 with synonym replacement for data augmentation and to 0.63 and 0.67 from data synthesizing by GPT4. Furthermore, the combination of data generation methods, including synonym replacement and GPT-4, and combining all methods display potential for improving overall classification performance compared to using a single method, as the major categories gain higher precision and the minor categories improve their recall.

***Table 2**. Average results of classification based on data generation techniques.*

	Micro Average			**Macro Average**		
	P	**R**	**F1**	**P**	**R**	**F1**
Baseline	0.63	0.75	0.68	0.75	0.56	0.64
SR	0.78	0.88	0.83	0.77	0.80	0.78
SfT	0.76	0.85	0.80	0.72	0.75	0.74
GPT	0.77	0.85	0.81	0.74	0.76	0.75
SR-SfT	0.80	0.90	0.85	0.80	0.82	0.81
SR-GPT	0.82	0.90	0.86	0.83	0.81	0.82
Sft-GPT	0.80	0.87	0.83	0.82	0.80	0.81
ALL	0.82	0.90	0.86	0.83	0.83	0.83

From Table 2, the micro average and macro average are calculated to display different aspects. Micro averaging calculates metrics globally by counting the total true positives, false negatives, and false positives across all categories as it considers each instance equally and calculates the metric over the aggregated counts. On the other hand, macro-averaging considers the metrics for each category independently and then calculates the average of them. The micro-average scores show that the best data generation solutions are both the combination of synonym replacement and GPT-4 and combining all methods for 0.82, 0.90, and 0.86 for precision, recall, and F1 score. In terms of macro-averaging, using the combination of all methods yields the highest F1 score of 0.83, and using both synonym replacement and GPT-4 comes in second with a 0.82 F1 score.To assess the statistical significance of the observed improvements, we performed a paired t-test comparing the F1-scores of the model with and without data augmentation across the 5 folds of cross-validation. The results showed a statistically significant improvement in F1-score with data augmentation ($t(4) = 3.25$, $p = 0.031$). The 95% confidence interval for the difference in F1-score was [0.02, 0.10], suggesting that the true improvement in F1-score is likely to be between 2 and 10 percentage points. The Cohen's d effect size was 1.45, indicating a large effect. These results provide strong evidence that data augmentation significantly improves the performance of the Thai text classification model.

4.2. Practical Results

In this part, we evaluate the use of the proposed method for classifying abstracts of the new projects. To achieve a reliable classification model, we trained all the collected data with the data generation using all methods as it gives the best classification result. The input is a Thai text which is a summary of a research proposal, and the model is tasked to classify the input into the 6 national strategy categories. In this experiment, there are 47 text instances for input, and the classification results are assessed by a research project owner. The assessment result can be one of the three options including 'Agree', 'Maybe', and 'Disagree'. The first and last options are as their literal meaning and will be given in case the classification result matches or does not match the category that a research project owner plans to assign, respectively. The 'Maybe' option will be given in the case of a research project owner who does not think of the category, but the assigned

category is also sensible. The assessment result is given in Table 3.

Table 3. *Assessment of classification results by category*

	NS	CE	DS	CJ	ED	RD	%
Agree	2	9	13	5	4	9	89.4
Maybe	0	1	2	0	0	0	6.38
Disagree	0	0	1	0	0	1	4.26

The results in Table 3 show that 89.4% of the classification matches the thoughts of the research project owner. There are three cases of unsure classifying results and two cases of disagreeing. The two ‘disagree’ cases are from the major categories, which are DS and RD, while the three 'maybe'' cases are from CE and DS. Upon analysis, we found that the terms, especially technical terms, in the text are mostly unknown for disagreeing cases; thus, the classification is voted to the categories containing the most similar untechnical terms instead. For the three unsure cases, project owners agree that the contents possibly belong to either group. Hence, this can give them options to choose the category for funding, which opens them to more alternatives in strategizing funding for the projects.

4.3. Discussion

This paper aims to solve the issue of imbalanced and insufficient data in Thai text classification. We apply data generation, including data augmentation and data synthesis, to improve data quantity and leverage the number of data instances per category. The generated data is then trained for a classification model in an ensemble fashion by combining the classification results of multiple models from different techniques to produce a final classification.The experiment results of classifying a small and imbalanced dataset of 251 documents with six categories indicate that using data generation improves the classification performance significantly in terms of precision, recall, and F1 score. For the single method, data augmentation by synonym replacement produces the highest F1 score of 0.83, which improves to a 0.15 score from the baseline. The combination of data generation methods, especially the combination of synonym replacement and GPT-4, slightly improves the performance of a single method for another 0.03 F1 score. This signifies that data generation is applicable to the task of multi-class Thai text classification with a small and imbalanced dataset and can solve the imbalanced data issue effectively. In terms of classification, the experiment results show that the model can appropriately classify unknown documents into the national strategy category by about 95%, while the remaining issue for incorrect classification results is the unknown technical terms that do not exist in the training data.The limitations of the proposed methods are mainly the linguistic resources for handling data generation. As Thai is a complex language with fewer public linguistic resources, it is difficult to apply all possible data generation techniques that require additional natural language processing tools and resources, such as a Thai sentence segmentation tool to accurately identify sentence and clause boundaries, a free-access annotated corpus for model tuning, and a lexical dictionary for word replacement. Thus, for data generation, this work is limited to synonym replacement for data augmentation and SMOTE for Text, and using GPT-4 from OpenAI for generating synthetic text data. Without open and reliable resources, the quality of the generated data will not be acceptable and may not uphold the expected performance in a classification task. Additionally, the small and imbalanced dataset is difficult to apply to the latest classification techniques, such as deep learning [20], convolutional neural network [21], BERT [22] since they require large amounts of data for training. By generating a large amount of data for the use of the latest techniques, the training data will be overwhelmed by synthetic data instead of the original data, which may lead to false classification and inappropriate use of AI. Hence, the issue of a proper number of synthesized data points used in text classification should be considered and is important to study further.

5. Conclusion

In this paper, we propose the use of data generation to solve the issue of text classification on a small and imbalanced dataset. We apply and compare the

applicable methods for Thai-based text, which are a synonym replacement method for data augmentation using a Thai synonym dictionary and a method for generating synthetic text data, including SMOTE for text, using GPT-4 from OpenAI. From the experiment results of classifying a small and imbalanced dataset of 251 documents, we found that the dataset with generated data noticeably improves the performance of classifying a Thai text in terms of precision, recall, and F1 score compared to the baseline. From the performance evaluation results, the overall F1 score increased as the major categories gained higher precision, and the minor categories had their recall greatly improved. For the single method, data augmentation by synonym replacement produces the highest F1 score of 0.83, which improves to a 0.15 score from the baseline. The combination of data generation methods, especially the combination of synonym replacement and GPT-4, slightly improves the performance of a single method for another 0.03 F1 score.

Author Statements:

- **Ethical approval:** The conducted research is not related to either human or animal use.
- **Conflict of interest:** The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper
- **Acknowledgement:**This project was financed by the Fundamental Fund year 2567 (FF67), allocated by the Thailand Science Research and Innovation (TSRI).
- **Author contributions:** The authors declare that they have equal right on this paper.
- **Funding information:** The authors declare that there is no funding to be acknowledged.
- **Data availability statement:** The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

References

[1] Thanajiranthorn, C., & Songram, P. (2020). Efficient Rule Generation for Associative Classification. *Algorithms 2020, Vol. 13, 299*, *13*(11), 299.https://doi.org/10.3390/A13110299

[2] Zhu, X., & Goldberg, A. B. (2009). Introduction to Semi-Supervised Learning. *Introduction to Semi-Supervised Learning*.https://doi.org/10.1007/978-3-031-01548-9

[3] Sun, S., Luo, C., & Chen, J. (2017). A review of natural language processing techniques for opinion mining systems. *Information Fusion*, *36*, 10–25. https://doi.org/10.1016/J.INFFUS.2016.10.004

[4] Batista, G. E. A. P. A., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter*, *6*(1), 20–29.https://doi.org/10.1145/1007730.1007735

[5] 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

[6] Arreerard, R., & Senivongse, T. (2018). Thai defamatory text classification on social media. *Proceedings - 2018 IEEE/ACIS 3rd International Conference on Big Data, Cloud Computing, Data Science and Engineering, BCD 2018*, 73–78. https://doi.org/10.1109/BCD2018.2018.00019

[7] Hemtanon, S., Phetkrachang, K., & Yangyuen, W. (2023). Classification and keyword extraction of online harassment text in Thai social network. *Bulletin of Electrical Engineering and Informatics*, *12*(6), 3837–3842. https://doi.org/10.11591/EEI.V12I6.5939

[8] Chumwatana, T. (2015). *Using sentiment analysis technique for analyzing Thai customer satisfaction from social media.* http://www.uum.edu.my

[9] Tanantong, T., Sanglerdsinlapachai, N., & Donkhampai, U. (2020). Sentiment. Classification on Thai Social Media Using a Domain-Specific Trained Lexicon. *17th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology, ECTI-CON 2020*, 580–583. https://doi.org/10.1109/ECTI-CON49241.2020.9158329

[10] Khamphakdee, N., & Seresangtakul, P. (2023). An Efficient Deep Learning for Thai Sentiment Analysis. *Data* *8*(5), 90. https://doi.org/10.3390/DATA8050090

[11] Klaithin, S., & Haruechaiyasak, C. (2016). Traffic information extraction and classification from Thai Twitter. *2016 13th International Joint Conference on Computer Science and Software Engineering, JCSSE 2016*. https://doi.org/10.1109/JCSSE.2016.7748851

[12] Wongsap, N., Lou, L., Jumun, S., Prapphan, T., Kongyoung, S., & Kaothanthong, N. (2018). Thai Clickbait Headline News Classification and its Characteristic. *2018 International Conference on Embedded Systems and Intelligent Technology & International Conference on Information and Communication Technology for Embedded Systems (ICESIT-ICICTES)*. https://doi.org/10.1109/ICESIT-ICICTES.2018.8442064

[13] Song, C., Xu, W., Wang, Z., Yu, S., Zeng, P., & Ju, Z. (2020). Analysis on the Impact of Data Augmentation on Target Recognition for UAV-Based Transmission Line Inspection. *Complexity*, *2020*(1), 3107450.https://doi.org/10.1155/2020/3107450

[14] Wei, J., & Zou, K. (2019). EDA: Easy Data Augmentation Techniques for Boosting

Performance on Text Classification Tasks. *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, 6382–6388. https://doi.org/10.18653/V1/D19-1670

[15] Jin, D., Jin, Z., Zhou, J. T., & Szolovits, P. (2019). Is BERT Really Robust? A Strong Baseline for Natural Language Attack on Text Classification and Entailment. *AAAI 2020 - 34th AAAI Conference on Artificial Intelligence*, 8018–8025. https://doi.org/10.1609/aaai.v34i05.6311

[16] Sennrich, R., Haddow, B., & Birch, A. (2015). Improving Neural Machine Translation Models with Monolingual Data. *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Long Papers*, *1*, 86–96.https://doi.org/10.18653/v1/p16-1009

[17]Kobayashi, S. (2018). ContextualAugmentation: Data Augmentation by Words with Paradigmatic Relations. *NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, *2*, 452–457. https://doi.org/10.18653/V1/N18-2072

[18] Ruangrajitpakor, T., Kingkaewkanthong, A., & Supnithi, T. (2018). Towards Electronic Version of the Royin Thai Dictionary from Information-Heavily Semi-structured Data Source. *Journal of Intelligent Informatics and Smart Technology*. https://ph05.tci-thaijo.org/index.php/JIIST/article/view/115

[19] Dietterich, T. G. (2000). Ensemble Methods in Machine Learning. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *1857 LNCS*, 1–15.https://doi.org/10.1007/3-540-45014-9_1

[20] Malakar, S., & Chiracharit, W. (2020). Thai Text Detection and Classification Using Convolutional Neural Network. *2020 59th Annual Conference of the Society of Instrument and Control Engineers of Japan, SICE 2020*, 99–102. https://doi.org/10.23919/SICE48898.2020.9240290

[21]Jitboonyapinit, C., Maneerat, P., & Chirawichitchai, N. (2023). Sentiment Analysis on Thai Social Media Using Convolutional Neural Networks and Long Short-Term Memory. *International Scientific Journal of Engineering and Technology (ISJET)*, *7*(1),74-80.https://ph02.tci-thaijo.org/index.php/isjet/article/view/246935

[22] Gatchalee, P., Waijanya, S., & Promrit, N. (2023). Thai text classification experiment using CNN and transformer models for timely-timeless content marketing. *ICIC Express Letter*, *17*(1), 91–101. https://doi.org/10.24507/ICICEL.17.01.91