ECTI Transactions on Computer and Information Technology
Journal homepage: https://ph01.tci-thaijo.org/index.php/ecticit/
Published by the ECTI Association, Thailand, ISSN: 2286-9131

Intent Mining of Thai Phone Call Text Using a Stacking Ensemble Classifier with GPT-3 Embeddings

Nattapong Sanchan[1]

ABSTRACT

Intent mining has recently attracted Natural Language Processing (NLP) research communities. Despite the extensive research on English and other widely spoken languages, intent mining in Thai remains unexplored. This paper proposes an extended framework for mining intentions in Thai phone call text. It utilized a stacking ensemble method with GPT-3 embeddings, constructed by systematically determining based and meta-classifiers using Q-statistic and F1 scores. Overall, the based classifiers consisting of Support Vector Classifier (SVC), k-nearest Neighbors (KNN), and Random Forest (RF) were derived with a meta-classifier, Logistic Regression (LR). We compared the mining results, derived through the proposed Stacking Ensemble Classifier (SEC), to 1) the individual base classifiers and 2) the three BERT baselines: BERT Multilingual Uncased, and BERT-th, and BERT Based EN-TH Cased. The results revealed that SEC could outperform SVC, KNN, RF, BERT Multilingual Uncased, and BERT-th, except BERT Based EN-TH Cased. However, a statistical analysis conducted using Friedman and Holm's post hoc tests reported no statistically significant difference between SEC and BERT Based EN-TH Cased, inferring that the two classifiers perform similarly practically.

Article information:

Keywords: Intent Mining, Intention Mining, Intent Classification, Intent Detection, Text Mining

Article history:
Received: September 8, 2024
Revised: November 21, 2024
Accepted: January 23, 2025
Published: January 31, 2025 (Online)

DOI: 10.37936/ecti-cit.2025191.258239

1. INTRODUCTION

Intent Mining (also known as Intention Mining and Intent Detection) has recently attracted Natural Language Processing (NLP) communities. It involves the process of extracting intentions expressed in text. Intention refers to a concept of goals, activities, or plans a user aims to do in the future [1-2]. By knowing and understanding the intentions of users, we are characteristically able to determine the proposition and the users' needs. Therefore, a system can deliver personalized content and appropriate recommendations to users. For instance, in a textual conversation, "My daughter is calcium deficient. What can I do?" indicates the intention of the speaker who needs calcium medication for his or her child [3]. By integrating intent mining into a system, it can recommend calcium-related products to serve the customer's needs. Another example is to determine intentions underlying a sequence of actions or processes performed by a user. When a user interacts in a system, performing a series of actions to finish a particular task, [4] aimed to observe the user's intentions underlying such actions instead of just describing the sequence of actions. Knowing users' intentions allows the system to improve the recommendation alternatives to users and the gaps between their intentions and the actions in the system. [5] exemplified an example of integrating intent mining in a movie review system. Their work enabled the understanding of reviewers' motivations, expectations, and emotions when recommending movies. Moreover, the analysis of intentions in the text also allows businesses to effectively and successfully launch marketing campaigns that align with customer's needs and desires [6], leading to critical decision-making and increased profits. Furthermore, in search engines, understanding users' intentions in searching keywords helps improve the accuracy and relevancy of search results and the efficiency of the information retrieval processes [7]. For these reasons, mining intentions in the text is exceptionally essential.

This paper aims to mine intentions in Thai phone call text. Understanding such intentions benefits

[1]The author is with the School of Information Technology and Innovation, Bangkok University, Pathum Thani 12120, Thailand , E-mail: nattapong.sa@bu.ac.th

business entities in various ways. For instance, classifying phone call purposes allows business entities to direct calls to relevant departments rapidly. A clear understanding of callers' intentions ensures callers can reach persons best suited to address their needs or concerns. Additionally, knowing the callers' objectives allows business entities to prepare adequate staff in each department to serve clients and organize appropriate staff training to assist clients effectively.

To mine intentions in Thai phone call text, we novelly explored the utilization of GPT-3 embeddings in classifying intentions with a stacking ensemble method. We systematically determined based classifiers and a meta-classifier using Q-statistics and F1 score. We explored Support Vector Classifier (SVC), XGBoost (XGB), Light Gradient Boosting Machine (LGBM), k-nearest Neighbors (KNN), and Random Forest (RF) in selecting based classifiers. We investigated the performance of the meta-classifiers with Gradient Boosting (GB) and Logistic Regression (LR) through the maximization of the F1 score. Overall, the based classifiers consisting of SVC, KNN, and RF classifiers were derived. The final meta-classifier is the LR classifier. In addition, we also investigated the performance and effectiveness of the individual base classifiers and a Stacking Ensemble Classifier (SEC) in Thai intent classification. We compared the results to the three BERT baselines used in [8]. The results revealed that SEC outperformed other competitive classifiers and was superior to two of the three BERT baselines. To conclude, the key contributions made in this paper are:

1) While existing research in Thai intent mining only focused on the BERT models, we leveraged the utilization of GPT-3 embeddings to mine intentions in Thai phone call text. This novelty sheds new light on applying a new type of text embeddings in Thai intent mining.

2) We extended the existing research by exploring the utilization of SEC and investigating its performance in mining Thai phone call text. We systematically defined the base and meta-classifiers using Q statistics and F1 scores, considering the balance between agreement in the correct and incorrect predictions.

3) Evaluating the proposed stacking ensemble classifier's performance outperforms other comparative classifiers and baselines.

2. RELATED WORKS

Researchers explored various approaches to mine intentions in text. One traditional approach was to use a rule-based method. [9] defined rules, keywords, and pattern matching to identify short query intentions in search engines and virtual assistants. Their systems were able to detect intentions based on specific keywords and the usage of regular expressions. Although this approach performed well in the experimented datasets, creating rules is laborious and time-consuming. New rules needed to be amended or defined for newly added complex-structured sentences.

Aside from a rule-based method, another well-known approach is a supervised learning method, which usually employs a classification technique for mining intentions in text. This technique trains and tests a system with a certain amount of labeled data. Generally, the classification evaluation reports precision, recall, F1 score, and with or without accuracy, which were calculated based on the correctness of predicted examples in the test dataset. Additionally, classification is commonly a backbone integrated with other approaches to mine intentions in text. For instance, a statistical approach was used in a work proposed by [10]. Textual representations of suicidal text were generated statistically using the Term Frequency-Inverse Document Frequency (TF-IDF) in the Bag of Words (BoW) model. Then, the textual representations were used in a classification task where Logistic regression, random forest, and XGBoost classifiers predicted suicidal cases. The results revealed that logistic regression outperformed other classifiers with an accuracy score of 89%.

Intent mining can also be viewed as an unsupervised learning task where a system mines intentions from a dataset without using labeled outputs. [11] explored an unsupervised learning approach to identify the intent of customer queries collected from customer service logs, E-mails, and chat transcripts. K-means and DBSCAN were used to cluster the text. Clusters with similar texts indicated related customers' intentions.

Moreover, a semi-supervised learning method can also be found in intent mining, where a system combines supervised and unsupervised approaches. [12] investigated a semi-supervised clustering that aimed to make the clustering results more valuable and applicable by integrating user feedback into the clustering process. A clustering of intention was performed before being incrementally refined with supervised feedback from users. Additionally, in the work proposed by [13], a graph-based semi-supervised was used for classifying intent tweets related to food and beverage, trips, employment and training, products and services, events and activities, and trifles. The proposed method generated an intent graph: the relationships between intent tweets and keywords. A set of labeled tweets was used as the inference for the intention class.

Additionally, related work illustrated the utilization of stacking ensemble methods in mining intentions in text. In this approach, a set of algorithms was ordered in a stack, and the output of one algorithm will be an input for the others. This approach has improved mining performance in predicting user purchase intentions [14]. [15] created a stacking of algorithms such as Gradient Boosting, Random Forest,

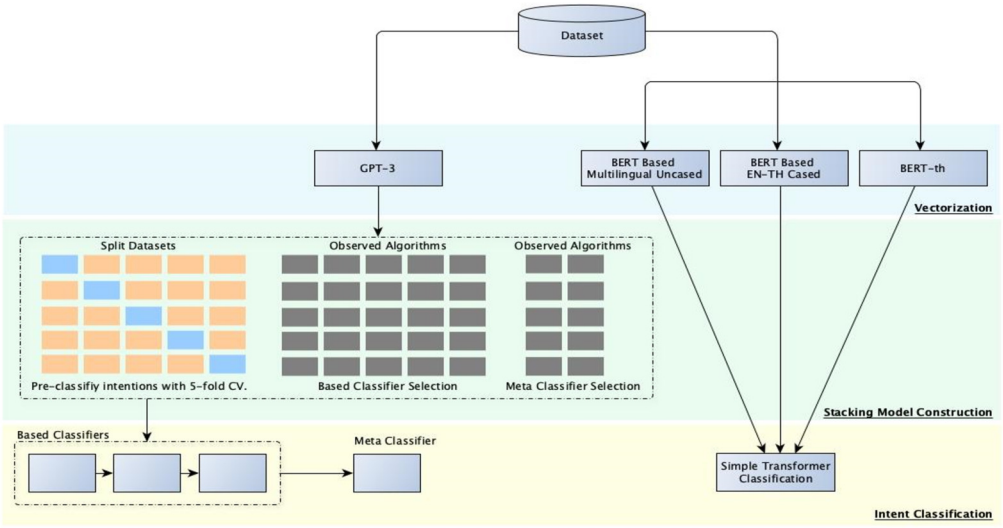

Fig.1: *A Proposed Framework for Thai Intent Mining Using a Stacking Ensemble Method with GPT-3 Embeddings.*

LightGBM, and Xgboost to identify marketing intentions. They proposed a method to combine models so that semantics features were derived. However, this work heavily relied on feature extraction. [16] also proposed a stacking method to mine potential users' intent on purchasing game items. Their central algorithms included Random Forest, XGBoost, SVM, and Random Forest. Their setting achieved a high F1 score of 90.71% but could only predict regular paying players, not the irrational spending players.

Furthermore, recent work has utilized transformer-based models to assist in intent mining. The models are intuitively grounded with deep learning techniques such as Recurrent Neural Networks (RNN) and Convolutional Neural Networks (CNN) [17]. One of the most popular models is BERT. The model was trained with a large amount of unlabeled text with the consideration of both left-to-right and right-to-left sides of textual sequences [18]. The generation of text embeddings in short messages [19], a fine-tuning BERT model to extend the original vocabularies [20], and a Label-Aware BERT Attention Network to assist a scoring of utterance in intentions [21] exemplify the application of BERT models to assist the intent mining. Another well-known transformer model is the Generative Pre-trained Transformer (GPT). It is a large language model with 175 billion parameters developed by OpenAI [22]. Examples of research that utilized GPT in intent mining are using ChatGPT with the setting of a zero-short environment to discover out-of-domain intentions [23] and using ChatGPT to improve intent classification [24].

Relevant related work has proposed a framework to mine intentions in Thai text related to gender equality and the law enforcement of same-sex mirages in Thailand [8]. They utilized variants of BERT models and generic classification algorithms to identify intentions in the dataset. Their results revealed that BERT Based EN-TH Cased outperforms other classifiers. However, their work only focused on utilizing individual classifiers.

Another relevant work is an example of using a stacking model in Thai text intent mining, published in a blogging platform [25]. The work only presented how a stacking method was created under a setting of TF-IDF and a random selection of stacked algorithms. Under this setting, no systematic approach was elaborated to select based and meta-classifiers.

Table 1: *Examples of Phone Call Texts in the Dataset.*

No.	Text	Classes
1	It cannot make calls. What is the reason?	Billing and payment
2	Can I apply for the free calling promotion, the TruveMove H?	Package promotion
3	AirmCard cannot connect to the Internet.	Internet
4	Is there a special price for data roaming?	International calling
5	I used to top-up TrueMoney, but I forgot the password.	True money
6	Yesterday my SIM card was lost. I will get a new SIM card in the Bangna-Trad area. Are there any shops around here?	Lost and stolen
7	I got an SMS, and the money was deducted. Moreover, can I move to the monthly payment?	Other queries

As the existing related work only focused on 1) Thai intention classification, with individual classifiers and BERT models, and 2) an unsystematic selection of stacked algorithms to mine intentions, this paper extended a framework presented in [8]. We leveraged the utilization of GPT-3 in classifying intentions with a stacking ensemble method. In addition, we systematically determined based classifiers and a meta-classifier using Q-statistics and F1 score. SVC, XGB, LGBM, KNN, and RF were explored when selecting based classifiers. Additionally, we investigated the performance of the meta-classifiers with GB and LR. Overall, the based classifiers consisting of SVC, KNN, and RF classifiers were derived. The final meta-classifier is the LR classifier. In addition, we also investigated the performance and effectiveness of the individual base classifiers and SEC in Thai intent classification. We conducted a hypothesis test to compare the results to the three BERT baselines used in [8]. The results revealed that the SEC outperformed other competitive classifiers and was superior to two of the three BERT baselines. Figure 1 illustrates the proposed framework, which will be elaborated on in the next section.

3. MATERIALS AND METHODS

3.1 Data Collection

We used TrueVoice's Mari dataset from [26] in our experiment. It contains Thai phone call text acquired from a mobile phone service provider in Thailand. The text expresses the callers' intentions, including *billing and payment, package promotions, Internet international calling, True money* (the topped-up money used for purchasing products and services provided by the True company), *lost and stolen*, and *other queries*. Table 1 illustrates the translated phone call text expressing users' intentions. For instance, the third example implies that a customer is seeking assistance regarding the Internet connection.

Initially, the dataset was divided into training and testing files, accounting for 12,939 and 3,236 records, respectively. Using cross-validation techniques, we combined the two files into one file to benefit the systematic selection of based and meta-classifiers in the stacking model construction and classification evaluation. There are 16,175 records, constituting 187,263 words in the dataset, leading to an average of 11.58 words per record. Table 2 illustrates the proportion of examples in each class.

3.2 Vectorization

In this process, Thai phone call text was transformed into vector representations (or text embeddings) so machine learning algorithms could process them later. We utilized GPT-3 and three BERT baselines for the vectorization, as shown in the sequential steps in Figure 1.

Table 2: *Proportion of Examples in Each Class.*

No.	Classes	Number of Examples
1	Billing and payment	5,984
2	Package promotion	3,729
3	Internet	2,477
4	International Calling	539
5	True money	307
6	Lost and stolen	353
7	Other queries	2,786

```
01  # Initialize OpenAI API
02  API_KEY = "PRIVATE API KEY"
03
04  # Specify an input sentence
05  sentence = "Air Card เชื่อม ต่อ อินเตอร์เน็ต ไม่ ได้"
06
07  # Submit a request to OpenAI API to
08  # acquire text embedding
09  req = OpenAI.Embedding.create(
10    model = "text-embedding-ada-002",
11    input = sentence)
12
13  # Extract the embedding from the API
14  embedding = req.data[0].embedding
15
16  # Save the embedding to file
17  np.save(file, np.array(embedding))
```

```
Example of GPT-3 Text Embedding:

[-0.04608885 -0.01602678 -0.00233865 ...
0.00755955 -0.00446054 -0.00982199]
```

Fig.2: *A Python Code Snippet for Acquiring GPT-3 Text Embedding.*

```
01  # Define packages
02  from transformers import AutoTokenizer,
03  AutoModel
04
05  # Specify a pre-trained model and
06  tokenizer
07    pmodel ="bert-base-multilingual-
08  uncased"
09    tokenizer =
10  AutoTokenizer.from_pretrained(pmodel,
11  use_fast=True)
12    model =
13  AutoModel.from_pretrained(pmodel)
14
15    # Specify text for tokenization
16    sentence = ["Air Card เชื่อม ต่อ อินเตอร์เน็ต ไม่
17  ได้"]
18
19    # Get Embedding
20    emb = tokenizer.batch_encode_plus
21  (sentence,  padding=True)
22
    print(emb)
```

```
Example of BERT Text Embedding:
{'input_ids': [[101, 11140, 18579, 100, 1030,
97004, 1043, 96993, 97007, 96991, 97004,
96999, 97007, 96993, 96991, 1048, 96997, 1048,
96990, 102]], 'token_type_ids': [[0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0]], 'attention_mask': [[1, 1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]]}
```

Fig.3: *A Python Code Snippet for Acquiring Text Embedding from a BERT Based EN-TH Cased Model.*

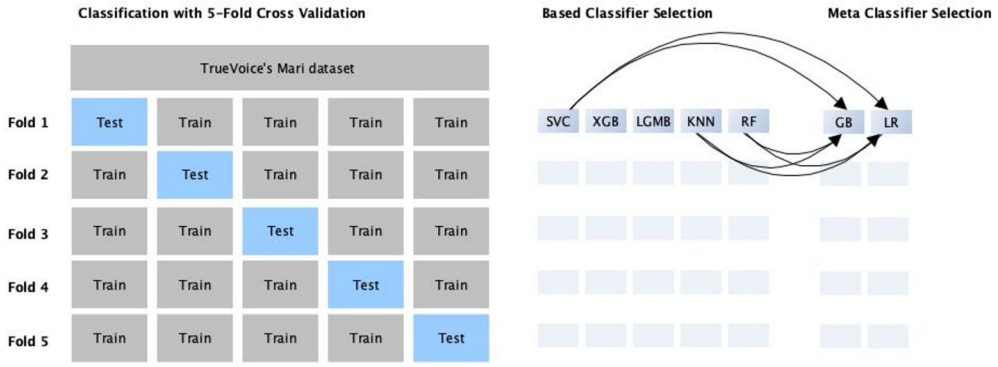

Fig.4: *The Construction of a Stacking Ensemble Method with a 5-fold Cross-validation Test.*

Table 3: *Micro-averaged F1 Scores of the Five Observed Classifiers.*

Classifiers	F1 Scores
SVC	0.7911
XGB	0.7849
LGBM	0.7732
KNN	0.7590
RF	0.6984

Table 4: *Pairwise Comparisons of Q-statistic among Five Observed Classifiers.*

	SVC	XGB	LGMB	KNN	RF
SVC	1.000	0.988	0.991	0.970	0.967
XGB	0.988	1.000	0.992	0.974	0.969
LGBM	0.991	0.992	1.000	0.969	0.969
KNN	0.970	0.974	0.969	1.000	0.959
RF	0.967	0.969	0.969	0.959	1.000

1) Generative Pre-trained Transformer 3 (GPT-3)
GPT-3 is the third version of the GPT series of large language models developed by OpenAI. The model was trained with many textual corpora collected from the Internet, books, and academic papers [22]. It was developed based on a transformer architecture, having self-attention mechanisms to efficiently process and understand natural language text with 175 billion parameters. We employed a model *text-tembedding-ada-002* specifically designed and optimized for generating text embedding. Figure 2 illustrates a Python code snippet for acquiring a GPT-3 embedding of a tokenized sentence "AirCard เชื่อม ต่อ อินเตอร์เน็ต ไม่ ได้" (AirCard cannot connect to the Internet.) and its result, having a dimensional size of 1536.

2) BERT
Bidirectional Encoder Representation from Transformers (BERT), developed by Google, is also a transformer-based model with self-attention mechanisms. It was bidirectionally trained with a large amount of text, such as Wikipedia text and books, meaning that the model considers context from both sides and each term in the sequence. The model can effectively capture the relation between terms and understand past and future contexts by analyzing the surrounding context in both directions. In the training process, BERT was trained to predict random words masked in sentences (Masked Language Model: MLM) and predict whether the subsequent sentences are logically related to the prior (Next Sentence Prediction: NSP). We employed three variants of BERT models specifically trained in Thai text, including BERT-based multilingual uncased, BERT-th, and BERT-based EN-TH cased, which were used as the baselines in this paper.

Figure 3 shows the Python code for acquiring text embedding from BERT Multilingual Uncased and its embedding. Each term in the input text is mapped to a unique ID of BERT's vocabulary, defined as *input_ids*. The numbers *101* and *102* correspond to $[CLS]$ and $[SEP]$, indicating the input sentence's starting and ending. The variable *attention_mask* indicates a value corresponding to each term, whether to process (1) or ignore (0).

3.3 Stacking Ensemble Classifier Construction

1) Overview of Explored Algorithms
To effectively construct the SEC, we reviewed recent literature on intent mining that utilized a stacking ensemble method and research on comparing classification algorithms [15] [27-28]. We explored a set of ro-

bustness algorithms that demonstrate notable results. For instance, a recent work explored the utilization of GB, RF, LGBM, and XGB as the base classifier. Additionally, LR was used as a meta-classifier [15], and GB was also assessed as it generally improves errors from their predecessors. We also explored SVC, which provides exceptional results in classification tasks [27-28] and Thai intent mining [8]. Note that the libraries from Scikit-learn with default configurations were employed to build the classifiers. The following section briefly discusses the algorithms that have been explored.

1.1) Support Vector Classification (SVC) was used to classify a multiclass of intentions. It aims to find a hyperplane that separates the examples into multiple classes.

1.2) Gradient Boosting (GB) was used to classify multiple intentions by sequentially constructing models that improve the errors of their predecessors.

1.3) XGBoost (XGB) was integrated to mine intentions by typically building an ensemble of decision trees using gradient boosting to improve the classification performance.

1.4) Light Gradient Boosting Machine (LGBM) generally constructs an ensemble of decision trees using a leaf-wise growth strategy to capture intentions.

1.5) K-nearest Neighbors (KNN) determines the proximity of data points for judging intention classes. New data are assigned to the nearest neighbor, and the intention class is predicted based on the majority of data points.

1.6) Random Forest (RF) is an ensemble classifier that constructs multiple decision trees. It combines the three results to justify the input's final intention class.

1.7) Logistic Regression (LR) generally estimates the probabilities of all classes in the dataset. Given an input, the class with the highest probability is assigned to the input.

2) Measuring Classifier Diversity

When constructing the SEC, based classifiers and meta-classifiers should be systematically defined. The selected base algorithms should be diverse and able to capture distinctive aspects of the dataset. To measure the diversity of the explored algorithms, we preclassified intentions in the Thai phone call text with a 5-fold cross-validation test, shown in Figure 4. The dataset was equally divided into five subsets - one is a test set, and the others are the training sets. Each algorithm is trained by the four split training sets and tested by the split test set in each fold. We reported the results using three evaluation metrics: precision, recall, and F1 score. Equation 1 to Equation 3 illustrates the equations of the evaluation metrics, calculated by substituting the number of true positives (TP), false positives (FP), and false negatives (FN) derived in the classification process into the equations. We aggregated the precision, recall, and F1 scores in each fold and reported the micro-averaged F1 score, as shown in Table 3. Precision indicates the proportion of examples that were accurately predicted by the classifier. Recall indicates, from all correct examples, how many correct examples are accurately captured. The harmonic mean of precision and recall, F1 score, is illustrated in Equation 3.

$$\text{Precision (P)} = \frac{TP}{TP + FP} \tag{1}$$

$$Recall\ (R) = \frac{TP}{TP + FN} \tag{2}$$

$$F1\ Score = \frac{2PR}{P + R} \tag{3}$$

Moreover, in each fold, we also measured the diversity of each algorithm by calculating Q-statistics values for each algorithm pair. The primary benefit of utilizing Q-statistics is the consideration of imbalanced classes in the dataset, providing the balanced measurement for true positive, true negative, false positive, and false negative examples. Equation 4 shows the Q-statistic formula, in which variable a indicates the number of correct examples in classifiers C_1 and C_2.b specifies the number of incorrect examples in C_1, but the correct ones appear in C_2.c and d are the opposite cases of b and a [29]. A Q-statistic value close to 1 indicates that pairwise classifiers are likely correlated and yield similar prediction results (low diversity). In contrast, a lower value close to 0 indicates no correlation between the classifiers, meaning that the classifiers predict different errors and thus are more diverse (high diversity).

$$Q = \frac{ad - bc}{ad + bc} \tag{4}$$

Table 4 illustrates the pairwise comparisons of Q-statistic among the five candidate classifiers. We selected the top three algorithms having the most diversity with the consideration of the F1 scores, shown in Table 3. From the tables, the assumption was that SVC has the highest F1 score with 0.7911, indicating the best-performing classifier and being a solid initial layer in the stacking model. The following selected algorithm was KNN. It has a moderate F1 score of 0.7590 but high diversity (low Q-statistic). It could help capture errors that SVC might miss. The final algorithm is RF. Despite its lowest F1 score, the algorithm has the highest diversity with SVC compared to other algorithms. RF could build a strong and stable final layer, reducing errors and enhancing the predictions of the stacking ensemble classifier.

3) Selecting Optimal Meta-Classifier

As shown in Figure 4, we investigated the performance of two standard algorithms, GB and LR, in selecting meta-classifiers. We derived two stacking

Table 5: *The F1 Scores Were Derived from the Classification of Intentions in Thai Phone Call Text.*

Folds	SVC	KNN	RF	SEC	BERT-th	BERT Based Multilingual Uncase	BERT Based EN-TH Cased
1	0.759	0.720	0.524	0.787	0.750	0.710	0.850
2	0.790	0.770	0.544	0.827	0.770	0.780	0.890
3	0.763	0.709	0.515	0.804	0.750	0.770	0.860
4	0.754	0.709	0.511	0.816	0.770	0.800	0.890
5	0.723	0.637	0.543	0.745	0.740	0.730	0.830
6	0.755	0.711	0.559	0.784	0.720	0.760	0.820
7	0.824	0.757	0.682	0.844	0.750	0.760	0.870
8	0.834	0.777	0.745	0.839	0.750	0.770	0.880
9	0.799	0.859	0.888	0.917	0.750	0.790	0.900
10	0.873	0.804	0.831	0.880	0.680	0.690	0.900
Micro-Average	0.787	0.745	0.634	0.824	0.743	0.756	**0.869**

Table 6: *P Values Were Derived from the Comparison between the Control Classifier (SEC) and Other Comparative Classifiers.*

Comparisons	‡	Unadjusted P	Adjusted P
SEC VS RF	4.554433	0.000005	**0.000032**
SEC VS KNN	3.519334	0.000433	**0.002163**
SEC VS BERT-th	3.415825	0.000636	**0.002544**
SEC VS BERT Multilingual Uncased	2.587746	0.009661	**0.028982**
SEC VS SVC	1.966687	0.049219	0.098439
SEC VS BERT Based EN-TH Cased	0.828079	0.407626	0.407626

ensemble classifiers: 1) (SVC, KNN, RF) → GB and 2) (SVC, KNN, RF) → LR. Following the previous classification pipeline, we select a meta-classifier that maximizes the F1 score. The results revealed that LR is a meta-classifier with the highest F1 score of 0.8062. In conclusion, we derived the SEC consisting of (SVC, KNN, RF) → LR.

3.4 Intent Mining

In this paper, we viewed the mining of intentions in Thai phone call text as a multi-classification task. As discussed in the previous section, we constructed SEC by selecting classifiers and meta-classifiers. The derived stacking classifier was utilized in the classification task. Finally, we obtained the results from the stacking classifier and the individual base classifiers. The results were compared to those of the three BERT baselines constructed using HuggingFace libraries. For the baselines, we employed a trailered classifier for BERT, the Simple Transformer Classification, with its default configuration. It is a transformer-based text classifier specifically used with BERT models.

4. EVALUATION AND DISCUSSION

4.1 Evaluation

The objective of the classification in this paper is to evaluate the performance of the proposed framework. We compared the classification results to other classifiers. Three BERT models, BERT-th, BERT Multilingual Uncased, and BERT Based EN-TH Cased were chosen as the baselines because they delivered strong performance in prior Thai intent classification research [8].

To assess the performance of the classifiers, we applied the Stratified 10-fold cross-validation to calculate the micro-averaged F1 scores of all classifiers. We aimed to report the F1 score because it is the harmonic mean of precision and recall. The results in Table 5 revealed that our proposed method, SEC, outperformed almost all the individual classifiers and baselines, except BERT-Based EN-TH Uncased. Compared to SVC, KNN, RF, BERT-th, and BERT Based Multilingual Uncased, SEC performed better by up to 4.70%, 10.60%, 29.97%, 10.90%, and 9.00%, respectively. For the BERT Based EN-TH Cased, SEC underperformed by 5.46%. Due to the slight difference in this underperformance, we further conducted a statistical analysis to investigate the statistical difference among the classifiers.

4.2 Statistical Analysis

To determine the significant differences among the four classifiers and the three baselines used in the Thai intent classification, we used the Friedman test [30] to analyze the F1 scores aggregated during the Stratified 10-fold cross-validation test [31], shown in Table 5. We utilized the Friedman test since we aimed to test multiple classifiers and did not assume

the variances' normal distribution and homogeneity [30]. We defined a null hypothesis (H_0) in the statistical test as no statistical difference exists among the classifiers. In addition, the alternative hypothesis (H_1) states a disparity among the comparison classifiers. We defined the control method as SEC to assess whether the proposed method was superior to other classifiers and not make pairwise comparisons [30].

By performing the classification on the TrueVoice's Mari dataset, the Friedman test revealed that the null hypothesis was rejected, meaning that there was a statistically significant difference among the classifiers, χ^2 (4) = 31.06, $p <= 0.00029$. Due to rejecting the null hypothesis, a post hoc test is required. We employed Holm's post hoc test by [32] to compare the F1 scores from the control classifier, SEC, against other classifiers and the baselines. Holm's post hoc test is crucial for controlling family-wise errors when testing multiple hypotheses [30]. The comparison results in Table 6, with p = 0.05, indicated statistically significant differences between the SEC and RF, KNN, BERT-th, and BERT Multilingual Uncased. However, there are no significant differences between 1) the SEC and SVC and 2) the SEC and BERT Based EN-TH.

4.3 Discussion

By comparing the classifiers with the application of GPT-3 embeddings, the micro-averaged F1 scores indicated that SEC outperformed other classifiers such as SVC, KNN, and RF. Additionally, by comparing the baselines integrated with BERT models, SEC outperformed two baselines: BERT-th and BERT Multilingual Uncased.

Comparing the baseline choice is also an important concern regarding textual languages. In the comparison, BERT-th (F1: 0.743) is an appropriate alternative when the dataset is purely Thai text, as the model was optimized for Thai-only data. Additionally, as some TrueVoice's Mari examples contain English terms, BERT Multilingual Uncased (F1: 0.756) is a better alternative as it supports multilingual datasets, yielding a higher F1 score than BERT-th. Lastly, among the three baselines, BERT-based EN-TH Cased (F1: 0.869) is the best alternative task for manipulating and analyzing Thai and English text when the highest F1 score was derived.

Compared to BERT Based EN-TH Cased, SEC had a slightly lower F1 score of 5.46%. A presumable factor making BERT Based EN-TH Cased superior is the quantity and variety of text used in its pre-training phase [33]. Unlike GPT-3, BERT models were trained bidirectionally, meaning that the models understand the context in sentences both right to left and left to right sides. GPT-3 was trained unidirectionally and could only predict the following terms by considering the previous ones. Additionally, BERT Based EN-TH Cased models were specially trained with abundant Thai text with the adjustment of vocabulary sizes and reduced unnecessary parameters [34], making the model superior to GPT-3.

Furthermore, the statistical analysis concluded that SEC could significantly outperform other comparative classifiers. The analysis revealed statistically significant differences among SEC and the three classifiers: SVC, KNN, and RF. However, the statistical analysis proved no difference between BERT Based EN-TH Cased and SEC. Nevertheless, this suggests that SEC is still a viable alternative, given its theoretical advantages in capturing aspects and handling complexities in Thai intent classification.

As a stacking ensemble method, each sequential algorithm in SEC has unique advantages crucial for Thai intent mining, implicating linguistic patterns and contextual variations. The stacking of SVC, KNN, RF, and the meta-model, LR, provides a powerful synergy for mining intentions in Thai text. Firstly, the ability of SVC to map input features into a high-dimensional space is enhanced with the utilization of GPT embeddings. SEC allows us to detect complex decision boundaries and differentiates closely related intentions such as " สอบถาม" (inquiry) and " ร้องเรียน " (complaint) where word ambiguity can arise due to the minor tone changes. Additionally, the advantage of KNN is the recognition of contextual variations where the general characteristic of Thai text is variant based on subtle linguistic patterns, dialect, levels of formality, politeness markers, and text tones. To illustrate, the intentions of two speakers might change to a slight variation due to their urgency expression: 1)"ช่วยเปิดสัญญาณโทรศัพท์ให้ด่วนค่ะ" (Please turn on the phone signal urgently) and 2) "ช่วยเปิดสัญญาณโทรศัพท์ให้หน่อยได้มั้ยคะ" (Could you please turn on the phone signal?). As the two sentences have closely related intentions, KNN groups the sentences together based on their feature proximity, providing effectiveness for the mining tasks where minor variations in wording or tone can indicate a shift in intentions. The following algorithm, RF, supports KNN in exploring integrations between features related to the aforementioned variations and accumulating multiple trees, reducing the risk of overfitting the rich information acquired from the high-dimensional GPT embeddings. Finally, after the base models have processed the text, LR serves as the meta-model, which combines outputs from each base model to produce the final prediction. LR captures the global trends and identifies patterns the base models might overlook. For this reason, given its theoretical advantages, utilizing a stacking ensemble method combines the strengths of multiple algorithms and, therefore, becomes essential in mining Thai intentions where the nature of the text is complex and variant.

While no statistical difference between BERT Based EN-TH Cased and SEC is found, SEC can surpass the baseline by adjusting parameters and se-

lecting diverse base classifiers and meta-models in future studies. Moreover, generalizing data in different domains might also be worth investigating in future research.

5. CONCLUSION

Since previous work only focused on using individual classifiers and BERT models to mine intentions in Thai text, this paper extends the previous work by systematically defining a stacking ensemble classifier in the mining process. We additionally leveraged the utilization of GPT-3 embedding in the classification task.

In our proposed model, the dataset was input into the system before their GPT-3 embeddings were generated in the vectorization process. Then, the embeddings were used as input for constructing SEC, which consisted of selecting the based classifiers and a meta-classifier. When selecting based classifiers, we explored SVC, XGB, LGBM, KNN, and RF. We also investigated the performance of the meta-models with GB and LR. By measuring the Q-statistic and F1 scores and maximizing the F1 scores of the overall classification task, we derived a stacking ensemble classifier as (SVC, KNN, RF) $\rightarrow$ LR. We compared the classification results from the stacking classifier against those of individual base classifiers. We also compared the results to the three BERT baselines. The results revealed that SEC could outperform SVC, KNN, RF, BERT Multilingual Uncased, and BERT-th, except BERT Based EN-TH Cased. However, a statistical analysis conducted using Friedman and Holm's post hoc tests reported no statistically significant difference between SEC and BERT Based EN-TH Cased, inferring that the two classifiers could be used interchangeably. For this reason, SEC is still a feasible option, given its theoretical advantages.

In future studies, research in Thai intent mining can be extended by exploring intent mining in different domains, adjusting parameters, and selecting diverse base and meta-models to increase the mining's performance. Additionally, future studies on applying a stacking ensemble classifier with BERT models could be beneficial, along with observing model fine-tuning. Furthermore, integrating an automatic summarization system to summarize intentions expressed in Thai text is also an alternative, allowing users to access content rapidly and digest intentions expressed in Thai text.

AUTHOR CONTRIBUTIONS

Nattapong Sanchan is the sole contributor to all aspects of this research, encompassing the research problems, methodology, software development, and evaluation. The author has reviewed and approved the final published version of the manuscript.

References

[1] R. C. Schank and R. P. Abelson, "Scripts, Plans, Goals, and Understanding," Hillsdale, New Jersey: Lawrence Erlbaum Associates, 1997.

[2] G. Khodabandelou, C. Hug, R. Deneckère and C. Salinesi, "Process Mining Versus Intention Mining," *Proceedings of BPMDS EMMSAD 2013, Enterprise, Business-Process and Information Systems Modeling*, pp. 466-480, 2013.

[3] J. W. Duan, Y. H. Chen, T. Liu and X. Ding, "Mining Intention-Related Products on Online Q&A Community," *Journal of Computer Science and Technology*, vol. 30, pp. 1054-1062, 2015.

[4] G. Khodabandelou, C. Hug, R. Deneckere and C. Salinesi, "Supervised Intentional Process Models Discovery Using Hidden Markov Models," *Proceedings of the IEEE 7th International Conference on Research Challenges in Information Science (RCIS)*, pp. 1–11, May 2013.

[5] V. D. Jadhav and S. N. Deshmukh, "Mining Movie Intention Using Bayes and Maximum Entropy Classifiers," *International Journal of Computer Applications*, vol. 975, no. 8887, pp. 8-19, 2016.

[6] M. Hamroun and M. S. Gouider, "A Survey on Intention Analysis: Successful Approaches and Open Challenges," *Journal of Intelligent Information Systems*, vol. 55, pp. 423-443, 2020.

[7] A. Rashid, M. S. Farooq, A. Abid, T. Umer, A. K. Bashir and Y. B. Zikria, "Social Media Intention Mining for Sustainable Information Systems: Categories, Taxonomy, Datasets and Challenges," *Complex Intelligent Systems*, vol. 9, pp. 2773–2799, 2021.

[8] N. Sanchan, "Mining Users' Intentions from Thai Tweets Using BERT Models," *Journal of Information Science and Technology*, vol. 13, no. 1, pp. 17-25, 2023.

[9] A. De and S. K. Kopparapu,"A Rule-Based Short Query Intent Identification System," *Proceedings of the 2010 International Conference on Signal and Image Processing*, pp. 212-216, 2010.

[10] O. A. Chidinma, S. Borah and R. Panigrahi, "Suicidal Intent Prediction Using Natural Language Processing (Bag of Words) Approach," *Proceedings of the International Conference on Computing and Communication (IC3 2020)*, pp. 147-153, 2021.

[11] H. D. Rebelo, L. A. de Oliveira, G. M. Almeida, C. A. Sotomayor, G. L. Rochocz and W. E. Melo, "Intent Identification in Unattended Customer Queries Using an Unsupervised Approach," *Journal of Information & Knowledge Management*, vol. 20, no. 3, pp. 1-26, 2021.

[12] G. Forman, H. Nachlieli and R. Keshet, "Clustering by Intent: A Semi-Supervised Method to Discover Relevant Clusters Incrementally," *Pro-*

ceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD 2015), pp. 20-36, 2015.

[13] J. Wang, G. Cong, X. Zhao and X. Li, "Mining User Intents in Twitter: A Semi-Supervised Approach to Inferring Intent Categories for Tweets," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 29, no. 1, 2015.

[14] J. Xu, J. Wang, Y. Tian, J. Yan and X. Li, "SE-stacking: Improving User Purchase Behavior Prediction by Information Fusion and Ensemble Learning," *PLOS ONE*, vol. 15, no. 11, pp. e0242629, 2020.

[15] Y. Wang, S. Liu, S. Li, J. Duan, Z. Hou, J. Yu and K. Ma, "Stacking-Based Ensemble Learning of Self-Media Data for Marketing Intention Detection," *Future Internet*, vol. 11, no. 7, p. 155, 2019.

[16] M. Li, L. Yang, W. Yixuan and Q. Zhang, "Predicting User Pay Conversion Intention Based on Stacking Ensemble Learning: Case Study of Free Value-Added Games," *Data Analysis and Knowledge Discovery*, vol. 8, no. 2, pp. 143–154, 2023.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser and I. Polosukhin, "Attention Is All You Need," *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 6000-6010, 2017.

[18] J. Devlin, M. W. Chang, K. Lee and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4171-4186, 2019.

[19] C. Oswald, S. E. Simon and A. Bhattacharya, "Spotspam: Intention Analysis-Driven SMS Spam Detection Using BERT Embeddings," *ACM Transactions on the Web (TWEB)*, vol. 16, no. 3, pp. 1-27, 2022.

[20] F. Fernández-Martínez, C. Luna-Jiménez, R. Kleinlein, D. Griol, Z. Callejas and J. M. Montero, "Fine-Tuning BERT Models for Intent Recognition Using a Frequency Cut-Off Strategy for Domain-Specific Vocabulary Extension," *Applied Sciences*, vol. 12, no. 3, p. 1610, 2022.

[21] T. W. Wu, R. Su and B. Juang, "A Label-Aware BERT Attention Network for Zero-Shot Multi-Intent Detection in Spoken Language Understanding," *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 4884-4896, 2021.

[22] T. Brown *et al.*, "Language Models Are Few-Shot Learners," *Proceedings of the 34th International Conference on Neural Information Processing Systems*, pp. 1811-1911, 2020.

[23] X. Song, K. He, P. Wang, G. Dong, Y. Mou, J. Wang, Y. Xian, X. Cai, and W. Xu, "Large Language Models Meet Open-World Intent Discovery and Recognition: An Evaluation of ChatGPT," *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 10291-10304, 2023.

[24] Z. Li, S. Larson, and K. Leach, "Generating Hard-Negative Out-of-Scope Data with ChatGPT for Intent Classification," *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 7634-7646, 2024.

[25] A. Daowraeng, "Text Classification in Thai Stacking Model," Accessed Nov. 21, 2024, [Online.] Available from: `https://medium.com/super-ai-engineer/การทำ-text-classification-ภาษาไทยด้วย-stacking-model-be12defb9-89e.`

[26] K. Viriyayudhakorn, "Truevoice-intent. GitHub," Accessed February 15, 2024, [Online.] Available from:`https://github.com/kobkrit/truevoice-intent.`

[27] P. Kaewnoo and T. Senivongse, "Identification of Software Problem Report Types Using Multiclass Classification," *Proceedings of the 2019 3rd International Conference on Software and e-Business*, pp. 104-109, 2019.

[28] K. Shah, H. Patel, D. Sanghvi and M. Shah, "A Comparative Analysis of Logistic Regression, Random Forest and KNN Models for the Text Classification," *Augmented Human Research*, vol. 5, no. 1, p. 12, 2020.

[29] L. I. Kuncheva and C. J. Whitaker, "Measures of Diversity in Classifier Ensembles and Their Relationship with the Ensemble Accuracy," *Machine Learning*, vol. 51, pp. 181–207, 2003.

[30] J. Demšar, "Statistical Comparisons of Classifiers over Multiple Data Sets," *The Journal of Machine Learning Research*, vol. 7, pp. 1-30, 2006.

[31] M. Abdel-Basset, H. Hawash, K. A. Alnowibet, A. W. Mohamed and K. M. Sallam, "Interpretable Deep Learning for Discriminating Pneumonia from Lung Ultrasounds," *Mathematics*, vol. 10, no. 21, p. 4153, 2022.

[32] S. Holm, "A Simple Sequentially Rejective Multiple Test Procedure," *Scandinavian Journal of Statistics*, vol. 6, no. 2, pp. 65-70, 1979.

[33] N. Sanchan, "Comparative Study on Automated Reference Summary Generation Using BERT Models and ROUGE Score Assessment," *Journal of Current Science and Technology*, vol. 14, no. 2, 2024.

[34] A. Abdaoui, C. Pradel and G. Sigel, "Load What You Need: Smaller Versions of Multilingual BERT," *Proceedings of SustaiNLP: Work-*

shop on Simple and Efficient Natural Language Processing, pp. 119-123, 2020.

Nattapong Sanchan is a faculty member of the School of Information Technology and Innovation at Bangkok University, Thailand. He graduated from the University of Sheffield, United Kingdom, in 2018, where he obtained a Ph.D. degree in Computer Science. Earlier, in 2011, he earned a Master of Information Technology in Data Management at Griffith University, Australia, and a Bachelor of Science in Computer Science at the International College, Burapha University, Thailand, in 2009. His research areas include opinion mining, automatic text summarization, information extraction, data mining, and natural language processing. He has been recognized for his work in these areas, receiving the Best Paper Award at the 18^{th} International Conference on Computational Linguistics and Intelligent Text Processing (CICLing) in 2017 for his paper titled "Gold Standard Online Debates Summaries and First Experiments Towards Automatic Summarization of Online Debate Data." In addition to his research, Nattapong Sanchan has been teaching at Bangkok University since July 2012. His teaching experience includes courses such as Data Analytics and Mining, Text Processing, Principles of Information Retrieval, and Database Systems.