CHAPTER 1

INTRODUCTION

1.1 Introduction and Problem Statement

The rise of e-commerce platforms like Shopee and Lazada has completely transformed how people in Thailand shop, leading to significant growth in the online market. These platforms offer convenience and a wide range of products. A key feature is the ability to read product reviews, which greatly influence purchasing decisions. Thai consumers trust fellow customers' opinions when evaluating product quality, features, and satisfaction. Reviews build trust in online shopping, reducing the risks of buying unseen items. Market reports confirm the significant impact of reviews on consumer behavior. As the Thai e-commerce market grows, reliance on product reviews for informed decision-making will continue, allowing Thai shoppers to confidently enjoy online shopping's benefits.

Among all the groups of products sold on these platforms, IT gadgets are particularly popular and highly demanded in the marketing industry. However, these IT gadgets can vary significantly in their usage characteristics, and they have specific quality indicators or descriptions that can pose a challenge for consumers. This variation makes it difficult for consumers to make informed purchase decisions based on the product descriptions provided. As a result, consumers who read reviews may need to spend a lot of time reading a large number of reviews to find good ones that can help them make purchase decisions, especially for consumers who are not knowledgeable in the IT product group. In addition, they may also receive unreliable reviews, which further results in buyers losing confidence in purchasing those products.

One solution to tackle time-consuming review reading is to use machine learning to sort reviews based on their helpfulness score. By doing so, readers can save time by avoid wasting time on irrelevant or unhelpful ones. However, the lack of a well-defined training dataset poses a significant hurdle in this work. Existing studies have limited information on how they collected and labeled the data, making it difficult to generalize their findings. To address this issue, a complete training dataset specifically designed for Thai language and IT gadget reviews is necessary. This dataset would greatly assist in creating and assessing precise machine learning models, giving consumers more confidence when making purchasing choices.

In this study, a contribution will be made by developing a machine learning-based system to predict the helpfulness of product reviews written in the Thai language specifically for IT gadgets. The aim is to utilize machine learning techniques to predict and sort these reviews based on their helpfulness score. To achieve this, the need of having a dedicated training dataset will be recognized. Therefore, another key contribution of this study will be the collection and creation of such a dataset comprising helpful product reviews in Thai.

1.2 Objectives

- To develop a machine learning model that effectively predicts helpfulness score of product reviews written in Thai for IT products on an E-commerce platform.
- To create a dataset of helpful product reviews in Thai Language for the IT product groups, with assigned helpfulness scores.

1.3 Significance of the Research

The findings of this study have broader implications for Thai language research. By addressing the challenge of predicting the helpfulness of product reviews in the context of IT gadgets and the Thai language, this research contributes to advancing our understanding of the Thai language. The development of a dedicated training dataset for Thai IT gadget reviews fills a significant gap in the existing literature and provides a valuable resource for future research in Thailand. The insights gained from this study can be leveraged to improve various applications, such as sentiment analysis and opinion mining, in industries where the Thai language is prominent. Ultimately, this research expands the scope of Thai language research and fosters advancements in language understanding and analysis.

1.4 Research Questions

- What machine learning model can effectively predict the helpfulness score of product reviews written in the Thai language for IT products on an e-commerce platform?
- How can a dataset of helpful product reviews in the Thai language for IT product groups be created, with assigned helpfulness scores?

1.5 Scope of the Study

The scope of this research study is to develop a machine learning-based system that predicts the helpfulness score of product reviews written in the Thai language specifically for IT gadgets. The study focuses on creating a dedicated dataset comprising helpful product reviews in Thai for the IT product groups, with assigned helpfulness scores. Data collection is conducted from the e-commerce platform Shopee, covering five popular categories of IT products. The dataset includes 1,200 reviews collected between January 1, 2022, and April 30, 2023.

1.6 Expected Benefits

- The development of a machine learning-based system to predict the helpfulness of product reviews aims to improve consumer decision-making, particularly for IT gadgets. By sorting and prioritizing reviews based on their helpfulness score, consumers can save time and gain confidence in their purchasing choices.
- Contributing to building trust in online shopping platforms, particularly in the Thai market, is the accurate identification and presentation of helpful product reviews. This can lead to increased customer satisfaction and loyalty.
- The creation of a dedicated dataset for Thai IT gadget reviews fills a gap in existing literature, providing a valuable resource for future research in Thailand. It expands the scope of Thai language research and contributes to advancements in language understanding and analysis.

1.7 Operation Definition

- Helpful product reviews: Reviews that provide valuable information, insights, or recommendations regarding IT gadgets. The helpfulness score of each review will be determined based on user ratings or feedback.
- Machine learning model: An algorithmic model that utilizes techniques such as natural language processing (NLP) and sentiment analysis to predict the helpfulness score of product reviews. Various machine learning algorithms, such as decision trees, support vector machines, or neural networks, will be explored and evaluated for their effectiveness.
- Dataset creation: The process of collecting a comprehensive and representative set of product reviews in the Thai language for IT gadget categories. The dataset will be manually labeled with helpfulness scores, indicating the usefulness of each review for potential consumers.

CHAPTER 2

LITERATURE REVIEWS

2.1 Product Review Helpfulness

Research on helping behavior provides a relevant foundation for studying the helpfulness of product reviews in the context of online shopping. Consumers, faced with limited time and resources, seek relevant information from a large volume of reviews to reduce uncertainty when making purchase decisions. Product reviews, whether from customers or experts, offer potential buyers' valuable insights into product features and usage experiences, facilitating the decision-making process. This provision of product reviews is considered a form of helping behavior, where reviewers demonstrate their desire to assist, commitment, and reciprocity in facilitating other consumers' purchase decisions.

product review helpfulness is defined as the degree to which consumers perceive a product review as capable of facilitating judgment or purchase decisions. Product review helpfulness comprises three dimensions: perceived source credibility, perceived content diagnosticity, and perceived vicarious expression. This definition draws upon Bach's research on helping behavior, which identifies three dimensions of helpfulness: trustworthy perception, problem-solving, and insight mediation. In the context of helping behavior and online shopping, consumers seek product reviews written by customers or experts as advisors to help them in their own purchase decisions. Thus, the three constructs of perceived source credibility, perceived content diagnosticity, and perceived vicarious expression align with the dimensions of helpfulness found in the helping behavior research field.

Perceived source credibility, akin to trustworthy perception, refers to advisees' perception of the advisors' trustworthiness in providing helpful behavior or information sincerely, rather than evasively pretending to be "warm-hearted." In the context of product reviews, this aligns with perceived source credibility, which relates to consumers' perception of the authors' credibility in providing accurate review information. When advisees perceive advisors as trustworthy, they maintain an open-minded attitude even when facing disagreements, enabling constructive resolution. Trustworthy advisors foster an environment where honest critique and annoyance can lead to deeper sharing of warm, positive feelings.

Perceived content diagnosticity, analogous to problem-solving, involves providing advice and intending to solve current problems. In helping behavior research, problem-solving occurs when advisors offer information to advisees, reinforcing the experience of receiving guidance for problem resolution. The conveyed information plays a crucial role in problem-solving performance. Unreliable information from advisors fails to help advisees distinguish interpretations and potential solutions, leaving the problem unsolved despite lengthy discussions. Perceived content diagnosticity aligns with problem-solving when consumers face the challenge of making shopping decisions in online purchases. It refers to the extent to which review information differentiates between alternative interpretations and solutions for a problem. Higher perceived diagnostic information enables more effective problem-solving. In the context of product reviews, high diagnostic information assists consumers in differentiating the benefits and concerns of a particular product, aiding them in their decision-making process.

Perceived vicarious expression, similar to insight mediation, involves gaining insights into others' functioning and understanding their inner world better, thereby comprehending their motivations for offering help. In line with insight mediation, perceived vicarious expression relates to the extent to which reviews convey vivid experiences of a product that readers can feel. This concept draws from vicarious experience in social learning theory. Reviews with high perceived vicarious expressions enable consumers to gain insights into why authors write reviews in a particular style, helping them understand the authors' perspectives when evaluating the target product. By reading product reviews with a high level of vicarious expression, consumers can learn about the usage experiences resulting from consuming the product. Thus, product reviews mediate authors' insights and assist consumers in their decision-making process based on authors' vivid expressions of their experiences.

In summary, the helpfulness of product reviews in online shopping can be understood through the dimensions of perceived source credibility, perceived content diagnosticity, and perceived vicarious expression. These dimensions align with the dimensions of helping behavior found in research, emphasizing trustworthiness, problem-solving, and insight mediation.

2.2 Prediction Model of Online Reviews

In this section, we will explore various predictive models that can be used for Thai text classification. Text classification is the process of automatically categorizing or assigning predefined labels to textual data based on its content, and it is a fundamental task in natural language processing (NLP) with applications such as sentiment analysis, spam detection, and topic classification. There are several commonly used predictive models for text classification, which can be broadly categorized into three types: traditional machine learning models, deep learning models and pre-trained models. These models are typically based on machine learning algorithms.

2.1.1 Traditional Machine Learning Models

Traditional machine learning algorithms or classical models are widely used for text classification tasks. These models often rely on handcrafted features extracted from the text, such as word frequencies, n-grams, or TF-IDF values.

Linear Regression: Linear regression is a powerful algorithm for text classification and regression tasks that involves predicting numerical values associated with text. While it is traditionally used for regression, it can be adapted for text classification by transforming the text features into numerical representations, such as using bag-of-words or word embeddings. The algorithm works by finding the relationship between these input text features and the target variable through fitting a linear equation to the data. The model then learns the coefficients for each feature, representing their impact on the target variable. To make predictions, the model multiplies these coefficients with the corresponding feature values and sums them up. By optimizing the coefficients using gradient descent, the model iteratively improves its ability to predict the target values accurately.

Support Vector Machines (SVM): SVM is a versatile algorithm widely used for text classification and regression tasks. In text classification, SVM takes the text features and maps them into a high-dimensional space, aiming to find the hyperplane that best separates the data points into distinct classes. This hyperplane maximizes the margin between the support vectors, which are the data points closest to the decision boundary. For regression, SVM seeks to find a hyperplane that approximates the relationship between the text features and the target variable. To accomplish this, SVM employs a kernel function to transform the data into a higher-dimensional space, where it can efficiently find the optimal hyperplane. The model is implemented using optimization algorithms like quadratic programming, which iteratively fine-tunes the hyperplane to minimize errors and maximize the margin.

Decision trees: Decision trees are highly adaptable algorithms commonly employed in text classification and regression tasks. In text classification, decision trees construct a hierarchical structure of nodes that make decisions based on the text features. The model efficiently splits the data by evaluating different features and their corresponding thresholds, recursively forming branches until it reaches leaf nodes containing the predicted class labels. Decision trees excel at capturing complex feature interactions and can handle both categorical and continuous text features. Moreover, they offer interpretability, as the decision-making process can be visualized as a flowchart-like structure. In text regression, decision trees partition the data in a similar manner but predict numerical values at the leaf nodes instead of class labels. By iteratively constructing the tree, optimizing for criteria such as information gain or Gini impurity, decision trees can accurately classify or predict values based on textual information. Their simplicity, interpretability, and ability to handle a wide range of data types make decision trees a valuable tool in text classification and regression problems.

K-Nearest Neighbors (KNN): KNN is a versatile non-parametric algorithm commonly used in text classification and regression tasks. In text classification, KNN assigns a class label to a text sample by comparing it with the k nearest neighbors in the feature space. It determines the most common class label among its neighbors, employing distance metrics like

Euclidean or cosine similarity to calculate the similarity between text samples. For text regression, KNN predicts the target variable by averaging the values of the k nearest neighbors. The algorithm's effectiveness hinges on choosing an appropriate value for k and selecting a suitable distance metric that captures the similarity between text samples accurately. KNN is relatively simple to implement and understand, making it accessible for various applications. However, it may face challenges in high-dimensional feature spaces and can be sensitive to noisy or irrelevant features.

2.1.2 Neural Network Machine Learning Models

A neural network model, in the context of text classification, refers to a computational model inspired by the structure and functioning of the human brain. It is a type of machine learning model that can analyze and process textual data by simulating interconnected artificial neurons, apply mathematical transformations and produce output signals. By adjusting the connections between nodes through a process called backpropagation, neural network models can learn to recognize patterns and relationships in the text, enabling them to make accurate predictions and categorize the input into predefined labels.

Deep Neural Networks (DNNs): DNN are models extensively employed in text classification and regression tasks. In text classification, DNNs leverage their multiple layers of interconnected neurons to learn intricate patterns and representations from the text features. Each layer applies nonlinear transformations to the input data, allowing the network to capture complex relationships. Through the process of backpropagation, the DNN learns the optimal weights for the connections between neurons, continually adjusting them to minimize the prediction error. This iterative learning process enables the network to adapt and refine its predictions over time. For text regression, DNNs can have a single output neuron to predict continuous values, leveraging the power of deep learning to approximate the relationship between the text features and the target variable.

Convolutional Neural Networks (CNNs): CNN are powerful deep learning models commonly used for image analysis, but they can also be adapted for text classification tasks. In text classification, CNNs leverage convolutional layers to scan the text features and extract local patterns and features. These convolutional layers use filters to identify specific patterns, such as n-grams or sequences of words, within the text data. By convolving these filters across the text, the network can capture relevant local dependencies and learn representations that are sensitive to specific textual features. The extracted features are then passed through fully connected layers for classification. During training, the CNN learns the weights of these filters through gradient-based optimization algorithms, adjusting them to minimize the classification error. This process allows the network to effectively identify and learn discriminative features from the text data

Recurrent Neural Networks (RNNs): RNN are specialized neural network models well-suited for handling sequential data, including text classification tasks. In RNNs, each word in the text sequence is processed sequentially, and the model maintains an internal memory state that allows it to capture dependencies and relationships between words and sentences. At each step, the hidden state of the RNN is updated based on the current input and the previous hidden state. This recurrent nature enables the network to retain information from previous words and consider context while processing subsequent words in the sequence. This capability is particularly valuable for understanding the semantic meaning and context of text. RNNs can be trained using backpropagation through time, which extends the traditional backpropagation algorithm to account for the sequential nature of the data. This allows the RNN to learn and adapt its internal memory representation to make accurate predictions for text classification or regression.

Long Short-Term Memory (LSTM): LSTM is a specialized type of Recurrent Neural Network (RNN) that effectively addresses the vanishing or exploding gradient problem inherent in traditional RNNs. LSTMs introduce a more sophisticated memory cell, equipped with gating mechanisms that regulate the flow of information within the model. These gating mechanisms, including the input gate, forget gate, and output gate, allow LSTMs to selectively retain or discard information based on the context and the input data. LSTMs can effectively capture long-term dependencies in text sequences, enabling them to understand the context and meaning of a sequence more accurately. This makes LSTMs particularly suitable for text classification and regression tasks that require a deep understanding of the sequence structure. LSTMs are implemented in deep learning frameworks, and their architecture has been widely adopted in natural language processing tasks, such as sentiment analysis, named entity recognition, and machine translation. The ability of LSTMs to handle sequential data and capture long-term dependencies has made them a powerful tool for solving text classification and regression problems, especially when dealing with complex textual data.

2.1.3 Pre-trained Machine Learning Models

A pre-trained model is a neural network model that has been trained on a large amount of data to learn general language patterns and understand text. Instead of building a model from scratch, pre-trained models take advantage of the knowledge acquired during this training process. These models are typically trained on tasks like language modeling or word prediction and develop an understanding of linguistic features and relationships. When used for text classification, pre-trained models offer several benefits. They save time and computational resources as they come with prior training, and their exposure to diverse data allows them to capture general language context and patterns. This is particularly advantageous for Thai text classification, considering the unique characteristics of the Thai language. To adapt the pre-trained model to a specific classification task, it is fine-tuned on a smaller dataset related to the target task, enabling it to specialize its knowledge.

Wangchanberta: Wangchanberta is a specific Thai language variant of the BERT (Bidirectional Encoder Representations from Transformers) model. It has been pre-trained on a large corpus of Thai text data using a transformer-based neural network architecture. Wangchanberta captures contextual information by considering the bidirectional relationships between words and their surrounding context. This model has been fine-tuned on various Thai text classification tasks, such as sentiment analysis and topic classification, allowing it to specialize in understanding the nuances of the Thai language. Wangchanberta has demonstrated impressive performance in Thai text classification and has become widely adopted in the Thai NLP community.

BERT (Bidirectional Encoder Representations from Transformers): BERT is a pre-trained model that has gained significant popularity in NLP. It has been trained on a vast amount of multilingual text data, including Thai. BERT leverages a transformer-based architecture to capture the contextual relationships between words. By considering the bidirectional information in the text, BERT generates contextualized word representations that encapsulate the meaning and semantics of the input. Fine-tuning BERT on Thai text classification tasks allows it to adapt its pre-trained knowledge to the specific nuances of the Thai language. BERT has achieved impressive results in various NLP benchmarks and is widely used for Thai text classification due to its ability to capture contextual information effectively.

RoBERTa: RoBERTa, which stands for A Robustly Optimized BERT Pretraining Approach is a variant of BERT that has been trained on a larger corpus of text data for a more extended period. RoBERTa improves upon BERT's pre-training methodology, enabling it to capture even more nuanced language patterns and semantic information. Like BERT, RoBERTa employs a transformer-based architecture to encode the contextual relationships between words. By fine-tuning RoBERTa on Thai text classification tasks, it can adapt its pre-trained knowledge to better understand the intricacies of the Thai language. RoBERTa has demonstrated exceptional performance in various NLP benchmarks and is considered a highly effective model for Thai text classification.

2.3 Data Cleaning and Data Preprocessing in Thai Language

Data cleaning and preprocessing in Thai language is essential for effective NLP tasks. Thai NLP poses unique challenges that highlight the need for thorough preprocessing. Firstly, Thai word segmentation is complex due to the absence of explicit spaces. Accurate segmentation is crucial for tasks like sentiment analysis and topic modeling. Informal language, slang, and non-standard spellings add further complexity. Handling these variations ensures accurate and reliable NLP analysis. Moreover, Thai text may include cultural references and idiomatic expressions requiring special consideration. Addressing these nuances avoids misinterpretations and biased analyses. Incorporating cultural lexicons and idiomatic expression handling improves the suitability of the dataset for meaningful Thai NLP analysis. To clean Thai language product reviews, the following steps are recommended.

2.2.1 Normalization

Normalization involves transforming data into a standardized and consistent form. This process includes converting the data to lowercase, removing punctuation, numbers, and special characters, expanding contractions, and correcting spelling errors. Normalization helps reduce noise and variability in the data, making it easier to compare and analyze.

2.2.2 Remove stop words

Stop words are commonly used words that contribute little to the overall semantic meaning of the text, such as articles, prepositions, and conjunctions. Removing stop words reduces the size and complexity of the data, allowing the focus to be on relevant and informative words for the given task. However, caution should be exercised not to remove words that may have contextual or domain-specific importance. For example, in question answering or dialogue systems, it may be necessary to retain stop words.

2.2.3 Tokenization

Tokenization is the process of dividing data into smaller units, such as words, sentences, or n-grams. Tokenization helps extract meaningful information and features from the data, preparing it for further analysis. The choice of tokenization level and method depends on the specific data and task. For instance, word-level tokenization is suitable for sentiment analysis, sentence-level tokenization is useful for summarization, and n-gram tokenization can be employed for topic modeling.

2.2.4 Vectorize data

Vectorization involves converting data into numerical representations that can be utilized by machine learning algorithms and models. Various methods and techniques are available for vectorizing data, including one-hot encoding, count vectorization, term frequency-inverse document frequency (TF-IDF), and word embeddings. The selection of the most appropriate method depends on capturing the features and semantics of the data and the specific task at hand. For instance, one-hot encoding is suitable for categorical data, count vectorization is commonly used for bag-of-words models, TF-IDF is useful for measuring document similarity, and word embeddings capture word similarity.

2.4 Related works

2.4.1 Helpfulness Prediction on Product Review

The prediction of the helpfulness of product reviews gained significant attention in the literature. Researchers have approached this task using supervised learning techniques, which involve training models on labeled data where the helpfulness of reviews is explicitly provided. They utilized various techniques to establish the correlation between review attributes and their level of helpfulness. For instance, (Zhang, Y. & Zhang D., 2014), and (Ngo-Ye, T. L. & Sinha A. P., 2014) developed a predictive model using SVM. (Goswami, K., Park, Y. & Song, C., 2017), and (Olatunji et.al., 2019) explored neural network algorithms for predicting product review helpfulness. (Kong, et.al., 2020) and (Sharma et.al., 2023) compare linear and non-linear models such as KNN, linear SVM, and neural network models such as DNN and CNN. (Bilal & Almazroi, 2022) used BERT a pre-trained model to predict helpful review. they pinpoint difficulty of requiring handcrafted or specialized pre-processing. Regarding the lack of literature on predictive helpful review in the Thai language context. However, there have been papers explored Thai NLP. (Bowornlertsutee & Paireekreng, 2022) conducted sentiment analysis on product reviews using four different models, including Linear model, LSTM, SVM, and SGD. (Manhem, et.al, 2020) used decision tree to perform sentiment analysis on hotel reviews. (Thetmueang & Jirawichitchai, 2017) employed four classical models to classify five level of sentiment in social reviews.

The review of existing literature on predicting the helpfulness of product reviews highlights three major types of models: classical models, neural network algorithms, and pre-trained models. These three types of models should be considered for model selection in the study.

2.4.2 Dataset Creation Methodology for Helpfulness Prediction

Various studies have explored the creation of datasets for predicting the helpfulness of product reviews. Factors that influence review helpfulness include credibility, content, and expression (Li, et.al., 2013). Content factors, such readability, number of votes, and depth, including expertise, and credibility, have also been identified in (Almutairi, et.al., 2019) and (Wu. J., 2017). Additionally, factors such as product quality, sentiment, uncertainty, and product category have been considered in (Siering & Muntermann, 2013) and (Barbosa, et.al., 2016). To determine the helpfulness, human annotation is commonly used to capture user opinions due to the subjectivity and individual perspectives of the product reviews (Singh, et.al., 2017). This process can be challenging due to its time-consuming and resource-intensive nature. (Snow et.al., 2008) and (Passonneau & Carpenter) propose efficient methods for annotating natural language processing (NLP) tasks by crowdsourcing with non-expert workers from platforms like Amazon Mechanical Turk. They demonstrate that using non-expert labels can be as effective as using annotations from experts. They also suggest using label scores instead of binary labels during training.

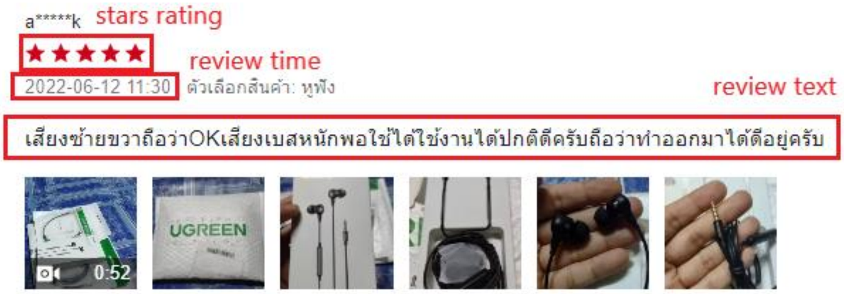
Currently, there is no study in create dataset in Thai for predict helpfulness in product review. To bridge this gaps, comprehensive and well-defined datasets are crucial. Our research aims to address these challenges by collecting and creating a dataset of helpful product reviews in the Thai language for IT products. This dataset will capture linguistic nuances and domain-specific characteristics, enabling accurate machine learning model to effectively predict and identify helpful product review.

CHAPTER 3

METHODOLOGY

3.1 Data Collection

The data collection methodology employed in this study utilized web scraping techniques to gather product reviews written in the Thai language from the renowned e-commerce platform, Shopee. The focus of the data collection was on five popular categories of IT products: smartwatches, keyboards and mice, speakers, earphones, and air purifiers. For each category, 240 reviews were collected, resulting in a total of 1,200 reviews. To ensure a fair representation of different opinions, 48 reviews were sampled from each set of 240 reviews, covering a range of ratings from 1 to 5 stars. This sampling strategy was designed to capture a diverse range of sentiments and viewpoints within the dataset. The scope of the data was explicitly limited to IT product reviews in the Thai language on Shopee. Data collection was specifically conducted from January 1, 2022, to April 30, 2023. The data scraped from each category was then filtered and pre-processed to collect the following fields: Review Time, Star Rating, and Review Text. These features are depicted in Figure 1.



a*****k

★★★★★

2022-06-12 11:30 ตัวเลือกสินค้า: หูฟัง

เสียงซ้ายขวาถือว่าOKเสียงเบสหนักพอใช้ได้ใช้งานได้ปกติดีครับถือว่าทำออกมาได้ดีอยู่ครับ

Figure 1. Data Collection from Shopee

3.2 Data Labeling

Data labeling methodology refers to the process of assigning meaningful tags or annotations to raw data. This crucial step involves human annotators who meticulously review and interpret the data, ensuring accurate labeling. In our study, we followed the methodology outlined by (Li, et.al., 2013). Each review was assessed based on three primary criteria: credibility, content, and expression. The credibility criterion focused on evaluating the trustworthiness, uncertainty, and reliability of the reviews. The content criterion aimed to assess the richness and depth of information provided in the reviews. Lastly, the expression criterion focused on assessing the mood and writing style of the reviewers. These criteria encompassed various aspects, including user sentiment, behavioral patterns, consistency. By considering these factors, we aimed to ensure a thorough analysis of the reviews.

Our annotation methodology involved a scoring system based on a scale of 0 to 3 for each criterion. Each annotator assessed the credibility, content, and expression of the reviews and assigned a score reflecting the extent to which the criteria were met. A score of 0 indicated that the criterion was not evident or poorly represented in the review, while a score of 3 indicated a clear and substantial representation of the criterion. To determine the overall helpfulness of each review, we summed the scores assigned to credibility, content, and expression. This summation score ranged from 0 to 9, representing the overall perceived helpfulness of the review. This helpfulness score served as the target variable for the helpfulness prediction task, which aimed to predict the level of helpfulness based on the annotated criteria.

$$helpfulness\ score = \sum_{i=1}^{n} \frac{(creditability + content + expression)_i}{number\ of\ annotator} \qquad (1)$$

3.3 Data Preprocessing

This section aims to provide the steps taken to clean and prepare the data for predicting the helpfulness of Thai reviews in the study. Data preprocessing is considered essential for enhancing the accuracy of the predictive model and addressing the challenges posed by the unique characteristics of the Thai language. The absence of word spaces and full stops, slang, multilingual language, including English words, and the presence of numbers and emoticons in reviews, making it difficult for machine learning models to accurately analyze the text. The data preprocessing techniques outlined in the paper (Khamphakdee & Seresangtakul, 2013) were followed to tackle these issues. The preprocessing steps involved symbol removal, number removal, English word removal, emoji and emoticon removal, text normalization, word tokenization, whitespace and tab removal, and single character removal.

3.4 Modeling

In this section, predictive models were developed to determine the helpfulness score of product reviews based on their informative content. Supervised learning techniques, specifically regression, were utilized to train and evaluate the models, establishing a relationship between the informative content of the reviews and their corresponding helpfulness scores. To compare different modeling approaches, three types of machine learning models were considered: classical models, neural network models, and pre-trained models. The inclusion of these three model types was informed by existing literature papers (Kong, et.al., 2020), (Sharma et.al., 2023), and (Bilal & Almazroi, 2022) that have employed these model types to predict review helpfulness.

3.4.1 Classical Models

We used four classical model's architectures, including Linear Regression (Galton ,1886), Support Vector Machine (Cortes & Vapnik, 1995), Decision Tree (Quinlan, 1986), and K Nearest Neighbors (Cover & Hart 1967). These models were trained using a regression approach to analyze the informative content of the reviews. We employed TF-IDF to convert the text into numerical features, allowing the models to establish patterns and relationships between these features and the helpfulness score. This approach enabled the prediction of helpfulness based on the informative content of the reviews.

3.4.2 Neural Network Models

We used four neural network architectures, including Deep Neural Networks (DNN) (Rumelart, et.al., 1986), Recurrent Neural Networks (RNN) (Elman, 1990), Convolutional Neural Networks (CNN) (Hochreiter & Schmidhuber, 1997), and Long Short-term Memory (LSTM) (Fukushima, 1980), in our study. These models were trained using a regression approach to analyse the informative content of the reviews. They excel at capturing complex relationships and abstract representations from the text data. In our implementation, we utilized a vocabulary-to-index mapping to establish a relationship between the words in the text and their corresponding indices. The neural network models typically consisted of an embedding layer, which converted the text into a dense vector representation. By adjusting the weights during the training process, the network learned the relationship between the review text and the helpfulness score. This enabled the models to make predictions based on the learned representations.

3.4.3 Pre-trained Models

We utilized four pre-trained models: WangchanBERTa (Lowphansirikul, 2021), BERT (Devlin, et.al., 2018), RoBERTa (Liu, et.al., 2019), and Thai-NER (Suriyachay, et.al., 2021). These models were developed by researchers in the field of natural language processing and initially trained by on a large-scale corpus of Thai language and, which included various textual sources such as social media posts, online reviews, news articles. The main objective of pre-training models was to capture language understanding for challenging aspects specific to the Thai language, including grammar, patterns, semantics, and contextual comprehension. To adapt these pre-trained models for our review helpfulness prediction task, we employed a process known as transfer learning. This involved fine-tuning the models using our labeled dataset, which consisted of review texts and their corresponding helpfulness scores. The fine-tuning process adjusted the internal weights and biases of the models to minimize the difference between the predicted helpfulness scores and the actual scores, enabling the models to make predictions of helpfulness scores.

3.5 Evaluation

This section presents methods to evaluate the performance of our system, including model performance evaluation and an assessment based on user experience.

3.5.1 Model Performance Evaluation

Model performance evaluation was conducted on the models with the aim to select the best model for our system by measuring the accuracy in predicting the helpful score of a review. The mean absolute error (MAE) was calculated to determine the average difference between the estimated score and the actual score of the review. This approach, presented by (Siering & Muntermann, 2013) indicates that a lower MAE indicates better alignment between the model's predictions and the actual helpfulness score. Furthermore, other metrics such as training time and validation time were also considered to support the process of model selection.

3.5.2 User Experience Evaluation

A user experience evaluation was conducted to assess the sorting of reviews based on their helpfulness, with the goal of measuring user experiences and satisfaction. The evaluation method included a comparison between the original review sorting method from the e-commerce platform and the implemented model for predicting helpfulness. Three criteria were included as user evaluation measures (Hu, 2009): perceived accuracy, user effort, and user loyalty.

Perceived accuracy: Perceived accuracy refers to the user's confidence in the accuracy of the product review sorting order in presenting the most valuable reviews. It assesses how users perceive the reviews to be sorted in a way that effectively identifies helpful reviews.

User effort: User effort measures the impact of the product review sorting order on the ease of locating the most relevant and informative reviews without much effort. It evaluates whether the implemented model for predicting helpfulness made it easier or harder for users to find the reviews they were looking for.

User satisfaction: User satisfaction evaluates how satisfied participants were with the product review sorting order in helping them make informed decisions. It aims to capture the overall satisfaction level of users with the sorting order and its effectiveness in assisting them in their decision-making process.

During the evaluation, participants were asked to rate their experiences based on three key questions:

- On a scale of 1 to 5 (1 = Not confident at all, 5 = Very confident), how confident are you in the accuracy of the product review sorting order in presenting the most valuable reviews?
- On a scale of 1 to 5 (1 = Made it much harder, 5 = Made it much easier), did the product review sorting order make it easier for you to locate the most relevant and informative reviews without much effort?
- On a scale of 1 to 5 (1 = Very dissatisfied, 5 = Very satisfied), how satisfied are you with the product review sorting order in helping you make informed decisions?

The intent behind the first question is to evaluate the perceived accuracy and assess how users perceive the reviews as tailored to their informative content, making it easier for them to identify helpful reviews. The second question aims to measure the subjective effort or time spent by users in completing the review sorting and decision-making process, and the final question measures user loyalty, indicating overall user satisfaction.

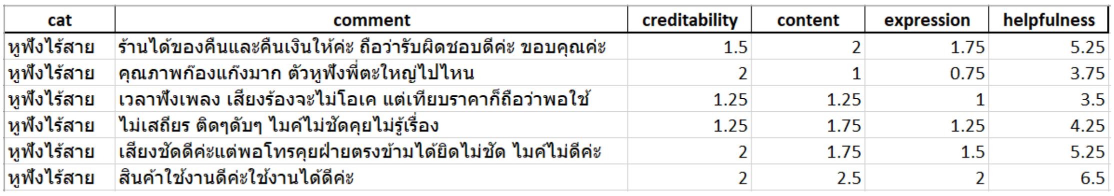
The final score for each criterion is calculated as the average score given by the participants for that specific criterion. For example, the final perceived accuracy score is the average score of all the participants' ratings on the perceived accuracy question. Similarly, the final user effort score is the average score of all the participants' ratings on the user effort question, and the final user satisfaction score is the average score of all the participants' ratings on the user satisfaction question. By calculating the average score, we can get an overall measure of the participants' perceptions and experiences regarding perceived accuracy, user effort, and user satisfaction. This allows us to summarize and compare the results of the user experience evaluation for the different criteria.

CHAPTER 4

RESULTS AND DISCUSSION

4.1 Data labeling Process

In the data labeling process, four annotators meticulously evaluated each of the 1,200 reviews based on three primary criteria: credibility, content, and expression. The annotators assigned scores to indicate the extent to which the criteria were met. These scores were then combined to calculate an overall helpfulness score for each review, consisting of Thai language reviews from Shopee across five IT product categories, was sampled to include 48 reviews from each category, covering a range of ratings. Out of the 1,200 reviews, 1,000 were used for training the model, while the remaining 200 were reserved for testing purposes.



cat	comment	creditability	content	expression	helpfulness
หูฟังไร้สาย	ร้านได้ของคืนและคืนเงินให้ค่ะ ถือว่ารับผิดชอบดีค่ะ ขอบคุณค่ะ	1.5	2	1.75	5.25
หูฟังไร้สาย	คุณภาพก๊องแก๊งมาก ตัวหูฟังพี่ตะใหญ่ไปไหน	2	1	0.75	3.75
หูฟังไร้สาย	เวลาฟังเพลง เสียงร้องจะไม่โอเค แต่เทียบราคาก็ถือว่าพอใช้	1.25	1.25	1	3.5
หูฟังไร้สาย	ไม่เสถียร ติดๆดับๆ ไมค์ไม่ชัดคุยไม่รู้เรื่อง	1.25	1.75	1.25	4.25
หูฟังไร้สาย	เสียงชัดดีค่ะแต่พอโทรคุยฝ่ายตรงข้ามได้ยิดไม่ชัด ไมค์ไม่ดีค่ะ	2	1.75	1.5	5.25
หูฟังไร้สาย	สินค้าใช้งานดีค่ะใช้งานได้ดีค่ะ	2	2.5	2	6.5

Figure 2 Helpfulness score calculating by each criterion

4.2 Model Selection

This section explains how we choose the most appropriate machine learning model for helpfulness prediction using k-fold cross-validation. The main criterion we considered for evaluation was the Mean Absolute Error (MAE). Additionally, we analyzed the train time and validation time of the models, which were obtained by averaging the durations across each cross-validation fold. Train time represents the average duration required to fully train the model until it reaches satisfactory or optimal performance, while validation time represents the average time taken by the model to process and generate predictions for each individual review in the validation data during cross-validation. The results of each model, obtained by performing 5-fold cross-validation, are presented in Table 1.

Table 1 Model Performance Evaluation using Cross-Validation for Model Selection

No.	Model	MAE	Train time (s)	Validation time (s)
	Classical Models			
1	Linear Regression	1.8838	0.46	0.000004
2	Support Vector Machine	1.0587	8.49	0.000152
3	Decision Tree	1.3793	1.95	0.000003
4	K-Nearest Neighbors	1.3119	1.67	0.000194
	Neural Network Models			
5	Deep Neural Network	1.6060	25.31	0.000062
6	Recurrent Neural Networks	1.6508	35.29	0.000121
7	Convolutional Neural Networks	1.1905	12.27	0.000068
8	Long Short-term Memory	1.6131	73.60	0.000656
	Pre-train Models			
9	WangchanBERTa	1.1974	2361.29	0.428230
10	BERT	1.2627	2048.74	0.265723
11	RoBERTa	1.0223	2483.60	0.338920
12	Thai-NER	1.1939	2680.10	0.297517

Among the evaluated models, the pre-trained model RoBERTa displayed the lowest MAE of 1.0223, indicating its superior accuracy in predicting the helpfulness score of IT gadget reviews compared to other models. This performance in terms of MAE led us to choose RoBERTa as the prediction model for this study. Although RoBERTa required a longer training time, it was deemed less significant considering that the training process was performed only once, and our primary objective was to identify the most accurate model. Additionally, after selecting the pre-trained model RoBERTa based on its superior performance in terms of MAE during cross-validation, we further evaluated its performance on a test set. The MAE obtained on the test set was 1.0447, which is very close to the MAE from the validation set (1.0223). This result demonstrates the generalization ability of RoBERTa to accurately predict the helpfulness score of unseen data. The RoBERTa model was initially trained on a large corpus of Thai Wikipedia texts. It also assigns UPOS tags to words in Thai sentences, which helps improve grammar and syntactic analysis. We can proceed with utilizing RoBERTa as the chosen model for predicting the helpfulness score of IT gadget reviews, as it consistently performs well on both validation and test data.

4.3 User Experience Evaluation

This section explains the user experience evaluation results conducted for the study. A total of 30 participants took part in the user study to evaluate two materials of the review sorting method. The first material is the original review order from Shopee (the baseline), and the second material is the review order based on our model, sorted from high helpfulness to low helpfulness scores. Each participant received two sets of materials, both materials used the same new unseen reviews, which consisted of 30 reviews for each category of the same product, resulting in a total of 150 reviews. Participants rated their satisfaction on a scale from 1 to 5 for the criteria mentioned in 4.5.2. The results of the user study are shown in Figure 2.

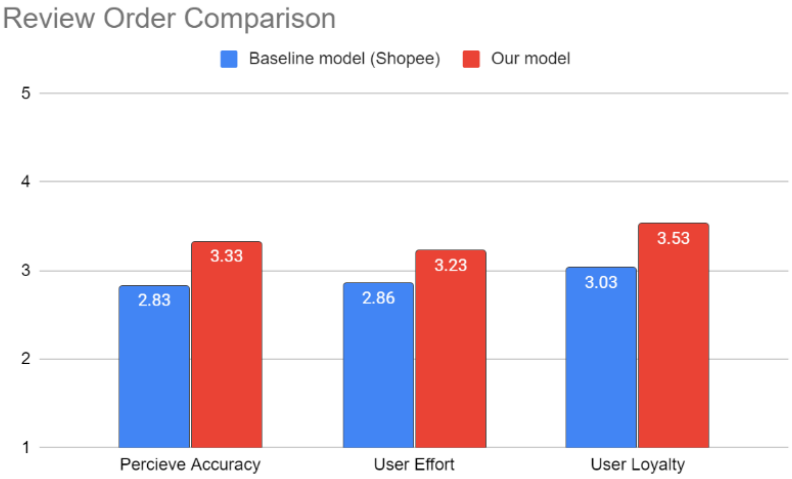


Figure 3 Comparing Satisfaction Scores of Review Order

According to the participant ratings, the reviews sorted by our implemented model have been perceived to exhibit several improvements compared to the baseline. Our model demonstrates higher accuracy in presenting reviews, as it received a rating of 3.33, surpassing the baseline model's rating of 2.83. Additionally, our model requires less user effort, as indicated by its rating of 3.23, outperforming the baseline rating of 2.86. Furthermore, our model fosters higher user loyalty, receiving a rating of 3.53, while the baseline obtained a rating of 3.03. These ratings highlight the enhanced performance and user experience provided by our model in terms of accuracy, user effort, and user loyalty, indicating its effectiveness in improving the review order on Shopee.

CHAPTER 5

CONCLUSION

This research aims to develop a machine learning system that predicts the helpfulness of product reviews in the Thai language for IT products. This will address the challenge of time-consuming review reading and provide consumers with a more efficient way to finding reliable and informative product reviews for informed purchasing decisions. We have defined two objectives for this study. First, we create a dataset of product reviews in Thai, specifically on IT product reviews with assigned helpfulness scores. Second, we utilize machine learning techniques and develop a system that accurately predicts and sorts reviews based on their helpfulness score. The methodology involved data collection through web scraping of Thai product reviews from Shopee. Data labeling was conducted by assessing credibility, content, and expression criteria, assigning scores to determine the overall helpfulness of each review. Data preprocessing techniques were applied to address the challenges posed by the characteristics of the Thai language in Thai Natural Language Processing and the handling of non-semantic characters. Three types of models: classical, neural network, and pre-trained models were developed and compared to predict the helpfulness scores of reviews based on informative content.

The results of our study demonstrate that the pre-trained model RoBERTa is a chosen model for predicting the helpfulness score of IT gadget reviews. It exhibited the lowest MAE during cross-validation, indicating superior accuracy compared to other models. Additionally, the generalization ability of RoBERTa was confirmed by its similar performance on the test set. This model selection process ensures reliable predictions for unseen data. Furthermore, the user experience evaluation revealed that the implemented model significantly improved the review sorting order on Shopee. Participants rated our system higher in terms of perceived accuracy, requiring less user effort, and fostering greater user loyalty compared to the Shopee baseline. These findings highlight the effectiveness of the implemented model in enhancing the user experience and assisting users in making informed decisions based on helpful reviews.

For future work, the scope of this research can be expanded to include other categories of products, allowing for the development of a multilingual system capable of handling diverse types of reviews. To enhance the accuracy and performance of the system, different model approaches, such as transformer-based models or ensemble techniques, can be explored. Furthermore, investigating the incorporation of user feedback and preferences into the model to personalize the review sorting process would be valuable.

REFERENCES

Zhang, Y., Zhang, D. (2014). Automatically predicting the helpfulness of online reviews. In Information Reuse and Integration (IRI), 2014 IEEE 15th International Conference on, IEEE.

Ngo-Ye, T. L., Sinha, A. P. (2014). The influence of reviewer engagement characteristics on online review helpfulness: A text regression model. Decision Support Systems, Vol. 61, pp. 47-58.

Goswami, K., Park, Y., Song, C. (2017). Impact of reviewer social interaction on online consumer review fraud detection. Journal of Big Data, Vol. 4, No. 1, pp. 15.

Olatunji, et.al. (2019). Context-aware helpfulness prediction for online product reviews. In Asia information retrieval symposium, Springer, pp. 56-65.

Kong, et.al. (2020). Predicting product review helpfulness: a hybrid method. IEEE Transactions on Services Computing.

Sharma, S. P., Singh, L., & Tiwari, R. (2023). Prediction of Customer Review's Helpfulness Based on Feature Engineering Driven Deep Learning Model. International Journal of Software Innovation, Vol.11, No.1

Bilal, M., & Almazroi, A. A. (2022). Effectiveness of Fine-tuned BERT Model in Classification of Helpful and Unhelpful Online Customer Reviews. Electronic Commerce Research.

Bowornlertsutee, P., Paireekreng, W. (2022). Emotion-based sentiment analysis model for categorizing online product review. Journal of Engineering and Digital Technology (JEDT), Vol. 10, No. 1.

Manhem, M., Dolah, S., Chunkaew, S., Mak-on, S. (2020). Model for the Classification of Feelings of Reviews Using Techniques Decision Trees: Case Studies Hotel Reservation Web Site. Journal of Science and Technology, Songkhla Rajabhat University, Vol. 2, No. 2.

Thetmueang, R., Jirawichitchai, N. (2017). Thai Sentiment Analysis of Product Review Online Using Support Vector Machine. Engineering Journal of Siam University, Vol. 18, No. 34.

Li, M., Huang, L., Tan, C.-H., Wei (2013). Helpfulness of online product reviews as seen by consumers: source and content features. International Journal of Electronic Commerce, Vol. 17, No. 4, pp. 101-136.

Almutairi, Y. A., Abdullah, M., Alahmadi, D. (2019). Review Helpfulness Prediction: Survey Periodicals of Engineering and Natural Sciences ISSN 2303-4521, Vol. 7, No. 1, June, pp. 420-432.

Wu, J. (2017). Review popularity and review helpfulness: A model for user review effectiveness. Decision Support Systems, Vol. 97, No. 6, pp. 92-103.

Siering, M., Muntermann, J. (2013). What Drives the Helpfulness of Online Product Reviews? From Stars to Facts and Emotions. In Proceedings of the 11th International Conference on Wirtschaftsinformatik, Leipzig, Germany.

Barbosa, J. L., Moura, R. S., Santos, R. L. d. S. (2016). Predicting Portuguese Steam Review Helpfulness Using Artificial Neural Networks. In Proceedings of the 22nd Brazilian Symposium on Multimedia and the Web, ACM.

Singh, J. P., Irani, S., Rana, N. P., et al. (2017). Predicting the 'helpfulness' of online consumer reviews. Journal of Business Research, Vol. 70, No. 6, pp. 346-355.

Snow, R., O'Connor, B., Jurafsky, D., Ng, A. Y. (2008). Cheap and Fast — But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks. Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, pp. 254-263, Honolulu.

Passonneau, R. J., Carpenter, B. (2014). The Benefits of a Model of Annotation. Transactions of the Association for Computational Linguistics, Vol. 2, pp. 311-326.

Khamphakdee, N., & Seresangtakul, P. (2021). Sentiment Analysis for Thai Language in Hotel Domain Using Machine Learning Algorithms. Acta Informatica Pragensia, Vol. 10, No. 2, pp. 155-171.

Galton, F. (1886). Regression Towards Mediocrity in Hereditary Stature. The Journal of the Anthropological Institute of Great Britain and Ireland, Vol. 15, pp. 246-263.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, Vol. 20, No. 3, pp. 273-297.

Quinlan, J. R. (1986). Induction of decision trees. Machine Learning, Vol. 1, No. 1, pp. 81-106.

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. IEEE Transactions on Information Theory, Vol. 13, No. 1, pp. 21-27.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. Nature, Vol. 323, No. 6088, pp. 533-536.

Elman, J. L. (1990). Finding structure in time. Cognitive science, Vol. 14, No. 2, pp. 179-211.

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biological cybernetics, Vol. 36, No. 4, pp. 193-202.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, Vol. 9, No. 8, pp. 1735-1780.

Lowphansirikul, L., et al. (2021). WANGCHANBERTa: Pretraining Transformer-Based Thai Language Models. arXiv preprint arXiv:2101.09635v2.

Devlin, J., et al. (2018). 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, Vol. 1, pp. 4171-4186.

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

Suriyachay, K., et al. (2021). Enhancement of Character-Level Representation in Bi-LSTM model for Thai NER. Science & Technology Asia, Vol. 26, No. 2, pp. 61-78.

Hu, R., Pu, P. (2009). A comparative user study on rating vs. personality quiz-based preference elicitation methods. Proceedings of the 2009 International Conference on Intelligent User Interfaces, Sanibel Island, Florida, USA, pp. 293-296.