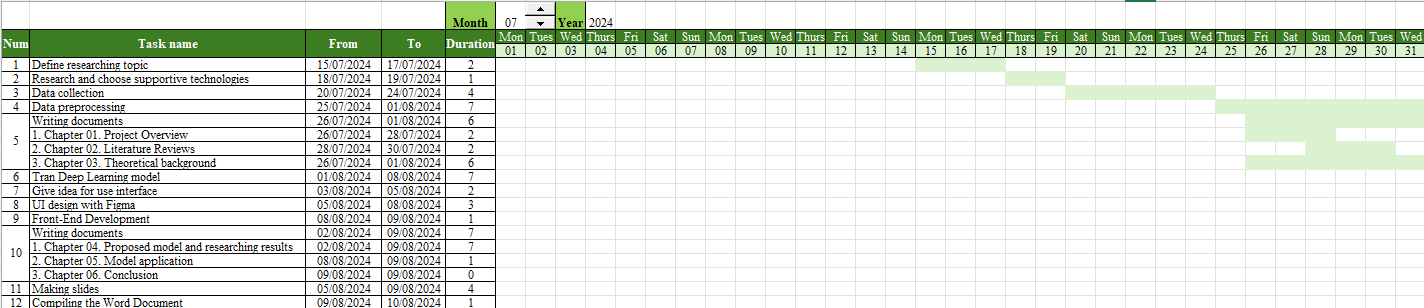
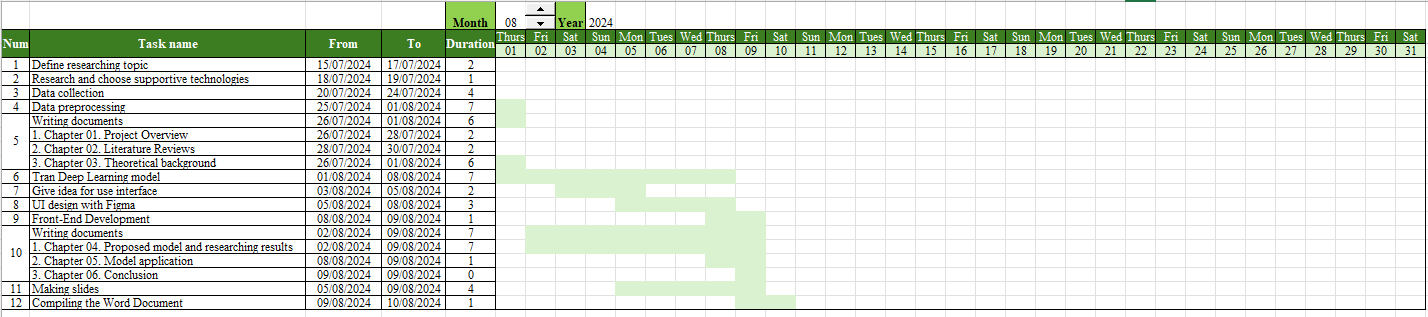
| **UNIVERSITY OF ECONOMICS AND LAW**  **FACULTY OF INFORMATION SYSTEMS**  **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**    **FINAL PROJECT**  **TOPIC: Applying Pho-BERT**  **in Customer Sentiment Analysis on Tiki Book Reviews**  **Subject: Deep Learning in Business Analytics**  **Lecturer: MSc. Nguyen Quang Phuc**  **Group 5**  **Ho Chi Minh City, August 10th, 2024** |
| --- |

# GROUP MEMBERS

| **No** | **Member** | **ID** | **Contribution** |
| --- | --- | --- | --- |
| 1 | Hoang Thi Nhat Quyen | K214162151 | 20% |
| 2 | Tran Thi Y Nhu | K214162149 | 20% |
| 3 | Ly Vu Thien An | K214161372 | 20% |
| 4 | Le Thi Bich Van | K214162159 | 20% |
| 5 | Le Tuan Nguyen | K214160993 | 20% |

# GANTT CHART

****

****

# ACKNOWLEDGEMENTS

We would like to extend our heartfelt gratitude to Master Nguyen Quang Phuc for your unwavering support, guidance, and dedication. Your expertise, knowledge, and dedication have been invaluable in our research and studies, serving as a constant source of motivation and confidence throughout our learning journey. We are deeply thankful for your enthusiastic mentorship and the time you have devoted to guiding us.

Despite our best efforts, we acknowledge that this project may still contain some imperfections. We sincerely welcome your feedback and suggestions to help us improve and refine our research.

**Group 5**

# COMMITMENT

Group 5.0 would like to commit to the research "Applying Pho-BERT in Customer Sentiment Analysis on Tiki Book Reviews" which is a research project carried out by the research group itself, under the guidance of Master Nguyen Quang Phuc.

Research processes and results were carried out and proposed by the group themselves. The information provided in this entire study is completely honest and is not copied from any other study. Referenced studies were clearly cited in the study.

We undertake to take full responsibility for our research if there are any errors.

**Group 5**

# TABLE OF CONTENTS

[GROUP MEMBERS i](#_heading=h.gjdgxs)

[GANTT CHART ii](#_heading=h.30j0zll)

[ACKNOWLEDGEMENTS iii](#_heading=h.1fob9te)

[COMMITMENT iv](#_heading=h.3znysh7)

[TABLE OF CONTENTS v](#_heading=h.2et92p0)

[LIST OF FIGURES vii](#_heading=h.tyjcwt)

[LIST OF TABLES viii](#_heading=h.3dy6vkm)

[LIST OF SYMBOLS AND ACRONYMS ix](#_heading=h.1t3h5sf)

[ABSTRACT x](#_heading=h.4d34og8)

[CHAPTER 1. OVERVIEW OF THE PROJECT 1](#_heading=h.17dp8vu)

[1.1. Reason for choosing the project 1](#_heading=h.3rdcrjn)

[1.2. Objectives of the study 2](#_heading=h.35nkun2)

[1.3. Scope of work 3](#_heading=h.1ksv4uv)

[1.3.1 Research Object 3](#_heading=h.44sinio)

[1.3.2. Research Scope 3](#_heading=h.2jxsxqh)

[1.4. Research Methods 3](#_heading=h.z337ya)

[1.5. Implementation process 3](#_heading=h.3j2qqm3)

[1.6. Report structure 4](#_heading=h.4i7ojhp)

[CHAPTER 2. LITERATURE REVIEW 6](#_heading=h.2xcytpi)

[2.1. Deep Learning 6](#_heading=h.1ci93xb)

[2.2. Customer sentiment analysis in e-commerce 7](#_heading=h.3whwml4)

[2.3. Textual Data Analysis 8](#_heading=h.2bn6wsx)

[2.4. Pho-Bert 9](#_heading=h.qsh70q)

[CHAPTER 3. THEORETICAL BACKGROUND 11](#_heading=h.3as4poj)

[3.1. Deep Learning 11](#_heading=h.1pxezwc)

[3.2. Bert 13](#_heading=h.147n2zr)

[3.3. Pho-Bert 16](#_heading=h.ihv636)

[3.4. Natural Language Processing (NLP) 18](#_heading=h.1hmsyys)

[3.5. Sentiment Analysis 20](#_heading=h.2grqrue)

[CHAPTER 4. PROPOSED MODEL AND EXPERIMENT 22](#_heading=h.3fwokq0)

[4.1. Problem and proposed model 22](#_heading=h.1v1yuxt)

[4.2. Data collection method 23](#_heading=h.2u6wntf)

[4.3. Data preprocessing 25](#_heading=h.28h4qwu)

[4.3.1. Preprocessing of assessment data 25](#_heading=h.nmf14n)

[4.3.2. Preprocess data across the entire dataset 29](#_heading=h.111kx3o)

[4.4. Topic Labeling 30](#_heading=h.4k668n3)

[4.5. Model Training 33](#_heading=h.sqyw64)

[CHAPTER 5. MODEL APPLICATION 37](#_heading=h.kgcv8k)

[5.1. User Interface 37](#_heading=h.34g0dwd)

[5.2.Interface implementation 39](#_heading=h.xvir7l)

[5.2.1. Home Page 39](#_heading=h.3hv69ve)

[5.2.2. Review page 41](#_heading=h.1baon6m)

[5.2.3. Write Review Page 42](#_heading=h.pkwqa1)

[Chapter 6: Conclusion 45](#_heading=h.1302m92)

[6.1. Conclusion 45](#_heading=h.3mzq4wv)

[6.2. Limit 45](#_heading=h.2250f4o)

[6.3. Development direction 46](#_heading=h.haapch)

[REFERENCES 47](#_heading=h.319y80a)

# LIST OF FIGURES

[Figure 1.1. Survey of customers on frequency of reading reviews [1] 2](#_heading=h.26in1rg)

[Figure 1.2. Average success scores of companies when conducting regular customer experience assessments [2] 2](#_heading=h.lnxbz9)

[Figure 1.3. Implementation process 4](#_heading=h.1y810tw)

[Figure 3.1. Deep neural network model [15] 11](#_heading=h.49x2ik5)

[Figure 3.2. How Machine Learning and Deep Learning work [16] 12](#_heading=h.2p2csry)

[Figure 3.3. BERT model with Masked LM [17] 14](#_heading=h.3o7alnk)

[Figure 3.4. Transformer model architecture [18] 15](#_heading=h.23ckvvd)

[Figure 3.5. PhoBERT model [19] 17](#_heading=h.32hioqz)

[Figure 3.6. Steps in the NLP process [20] 19](#_heading=h.41mghml)

[Figure 3.7. Commonly used machine learning and deep learning models [9] 20](#_heading=h.vx1227)

[Figure 4.1. Proposed model (proposed by the research team) 22](#_heading=h.4f1mdlm)

[Figure 4.2. Data after preprocessing 28](#_heading=h.2lwamvv)

[Figure 4.3. Number of data rows in each set 28](#_heading=h.3l18frh)

[Figure 4.4. Dictionary after standardization 29](#_heading=h.206ipza)

[Figure 4.5. Dictionary of dataset 31](#_heading=h.1egqt2p)

[Figure 4.6. Create and fit TF-IDF transformer 31](#_heading=h.3ygebqi)

[Figure 4.7. Some lines of data after TF-IDF calculation 32](#_heading=h.2dlolyb)

[Figure 4.8. Initialize training components and learning schedule 33](#_heading=h.3cqmetx)

[Figure 4.9. Text Tokenization Processing 33](#_heading=h.1rvwp1q)

[Figure 4.10. TF-IDF feature conversion process to tensor 34](#_heading=h.4bvk7pj)

[Figure 4.11. Prediction and loss calculation 34](#_heading=h.2r0uhxc)

[Figure 4.12. Calculate correct predictions and update weights 34](#_heading=h.1664s55)

[Figure 4.13. Result of training model 35](#_heading=h.3q5sasy)

[Figure 4.14. Test and predict model with example text 35](#_heading=h.25b2l0r)

[Figure 5.1. Bookstore homepage 36](#_heading=h.1jlao46)

[Figure 5.2. Book detail page 37](#_heading=h.43ky6rz)

[Figure 5.3. Book review writing page 37](#_heading=h.2iq8gzs)

[Figure 5.4. Home Page 38](#_heading=h.1x0gk37)

[Figure 5.5. Search bar ở Home Page 39](#_heading=h.4h042r0)

[Figure 5.6. Change the product when clicking the Next or Previous button 39](#_heading=h.2w5ecyt)

[Figure 5.7. Review Page 40](#_heading=h.3vac5uf)

[Figure 5.8. Search functionality on the Review page 41](#_heading=h.2afmg28)

[Figure 5.9. Write Review Page 41](#_heading=h.39kk8xu)

[Figure 5.10. Message Box after the user submits their comment (Positive Message Box) 42](#_heading=h.1opuj5n)

[Figure 5.11. MessageBox Neutral 43](#_heading=h.48pi1tg)

[Figure 5.12. MessageBox Negative 43](#_heading=h.2nusc19)

# LIST OF TABLES

[Table 4.1. Tiki Books Dataset description 23](#_heading=h.3tbugp1)

[Table 4.2. Definition of abbreviations and slang words 25](#_heading=h.37m2jsg)

[Table 4.3. Definition of vowels 26](#_heading=h.1mrcu09)

[Table 4.4. Name of Punctuation characters 27](#_heading=h.46r0co2)

[Table 4.5. Topic label definitions 30](#_heading=h.2zbgiuw)

# LIST OF SYMBOLS AND ACRONYMS

| **Word** | **Definition** |
| --- | --- |
| UI | User Interface |
| TF-IDF | Term Frequency Inverse Document Frequency |
| BERT | Bidirectional Encoder Representations from Transformers |
| PGD | Projected Gradient Descent |
| B2C | Business-to-consumer |
| NLP | Natural Language Processing |
| NLU | Natural language understanding |
| SVM | Support Vector Machine |
| CNN | Convolutional Neural Network |
| LSTM | Long Short-Term Memory |
| EDA | Exploratory Data Analysis |
| HSD | Hate Speech Detection |
| AI | Artificial Intelligence |
| MLM | Masked language modeling |
| NSP | Next Sentence Prediction |
| NER | Named Entity Recognition |
| re | Regular Expression |
| NFC | Normalization Form Composed |
| CPU | Central processing unit |
| GPU | Graphics processing unit |

# ABSTRACT

In the context of rapidly evolving e-commerce, the ability to understand and enhance customer experience has emerged as a critical determinant of business success. This project focuses on analyzing customer reviews on Tiki, one of the leading e-commerce platforms in Vietnam. The primary objective of this study is to leverage comment data to conduct sentiment analysis, thereby offering valuable insights that can assist Tiki in optimizing its products and services, ultimately leading to increased customer satisfaction and loyalty.

To achieve this, we employ the deep learning model PhoBERT, which is an optimized version of BERT specifically designed for the Vietnamese language. This model enables us to accurately identify and classify the emotions expressed in customer comments, as well as detect underlying trends in feedback. By analyzing these sentiments, we aim to uncover significant patterns that reflect consumer experiences, allowing Tiki to make informed adjustments to its business strategies. The application of PhoBERT not only enhances the accuracy of sentiment analysis but also offers efficiency in processing large volumes of data compared to traditional methods.

The user interface developed for this project is designed with simplicity and intuitiveness in mind, allowing users to easily input their comments and receive immediate sentiment analysis results. This user-friendly design aims to provide a seamless experience, encouraging consumers to actively engage in the feedback process. By facilitating easy access to sentiment insights, Tiki can better understand customer perceptions and preferences, fostering a more interactive relationship with its user base.

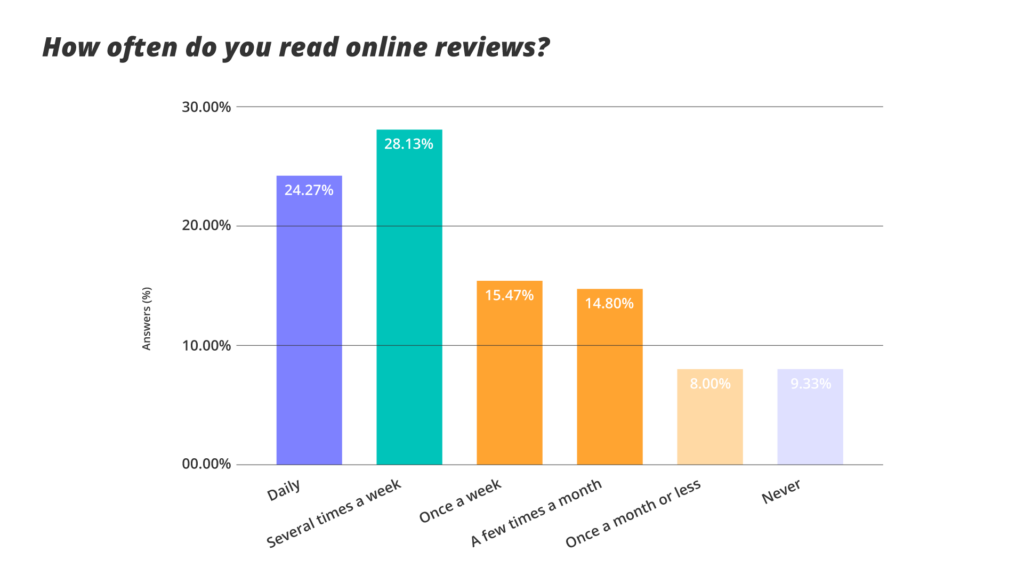
The findings of this research not only provide Tiki with actionable insights to improve its services but also pave the way for future studies in the realm of e-commerce. The insights gained from this project can contribute to a broader understanding of customer sentiment trends, promoting innovation and sustainable development within the industry. Overall, this study underscores the importance of leveraging advanced analytical techniques to enhance the customer experience in an increasingly competitive market.

**Keywords:** *Customer sentiment analysis, Deep learning, Natural language processing, e-commerce.*

# CHAPTER 1. OVERVIEW OF THE PROJECT

Chapter 1 provides an overview of e-commerce and the crucial role of customer experience in today’s competitive environment. This chapter introduces the Tiki platform, analyzes the factors influencing customer satisfaction, and clearly outlines the research objectives. Additionally, it discusses the research methods, processes, and tools that will be used, laying the foundation for the subsequent chapters.

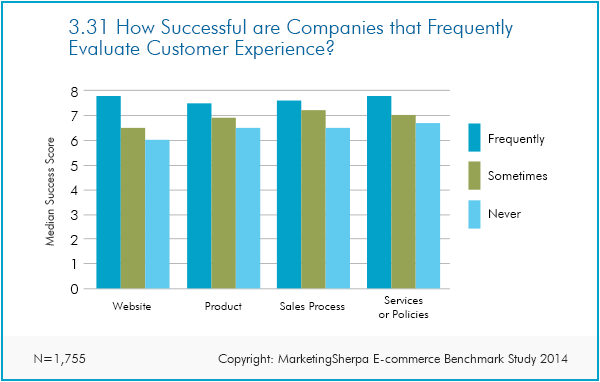
## 1.1. Reason for choosing the project



#### Figure 1.1. Survey of customers on frequency of reading reviews [1]

In today's competitive economy, whichever business or organization captures the majority of consumer tastes, that business or organization will certainly develop. Customer experience with a business organization plays an important role in influencing the decision to consume, recommend or return to that organization. Positive reviews will help businesses recognize the factors that create a positive customer experience when experiencing products and services, promoting loyalty and retaining customers longer. As for negative reviews, businesses promptly grasp the bad factors that negatively affect sales, leading to word of mouth negative information about the business. Especially in today's digital age, customers always need to research and understand the quality of products and services they want to consume from any business. Consumers have access to a wealth of information and are able to research products or services online and read reviews before making a purchase. According to a survey by the Pew Research Center’s Internet & American Life Project, about 58% of U.S. adults will conduct product and service research on the Internet before deciding to buy, and about 24% of customers have posted comments and reviews online after experiencing a product or service.

The chart above shows how often 750 people in Germany, Australia, Switzerland, France, and Italy read online reviews, showing that the majority of respondents read online reviews daily or several times a week. Once you’ve identified online reviews as having a significant impact, it’s important to collect them regularly.



#### Figure 1.2. Average success scores of companies when conducting regular customer experience assessments [2]

The bar chart of the success of companies included in the Marketingsherpa Ecommerce Benchmark Study shows that companies that regularly evaluate customer experience have higher success scores than companies that only occasionally or never evaluate customer experience.

In the context of increasingly growing and fiercely competitive e-commerce, analyzing customer reviews on the Tiki platform has become an important research topic. These comments provide valuable information not only about product quality but also about the overall shopping experience on Tiki. Through analyzing review data, Tiki can improve products and services, optimize marketing strategies, and manage customer relationships more effectively.

Therefore, this study was chosen to exploit the great potential of comment data in improving performance and enhancing Tiki's competitiveness in the e-commerce market.

## 1.2. Objectives of the study

The research was conducted to achieve the following main objectives:

* **Objective 1:** Train a model using deep learning methods to analyze customer sentiment with input data being customer reviews of products on the Tiki e-commerce platform.
* **Objective 2:** Design and build a set of interactive user interfaces, helping users grasp information about product reviews from historical data.

Based on the specific objectives set by the research team from the beginning, the process of analyzing, training the model and building the interactive interface will become more oriented.

## 1.3. Scope of work

### 1.3.1 Research Object

The object used by the research team to analyze customer sentiment is the customer review dataset of books on the famous e-commerce platform in Vietnam, Tiki.

### 1.3.2. Research Scope

The research analyzes the reviews collected and synthesized in the customer review dataset.

## 1.4. Research Methods

The data used by the research team for analysis is the customer review dataset of book products on the Tiki e-commerce platform. The research aims to analyze customer sentiment and build a deep learning model for this subject.

* **Method 1:** Collect secondary data on customer reviews from the Tiki platform, identify and select the necessary attributes for analysis to achieve the research goal of customer sentiment.
* **Method 2:** Build a data processing process, build a deep learning model to analyze customer sentiment in Vietnamese using PhoBERT, thereby discovering topics, opinions and satisfaction levels from the responses.

## 1.5. Implementation process

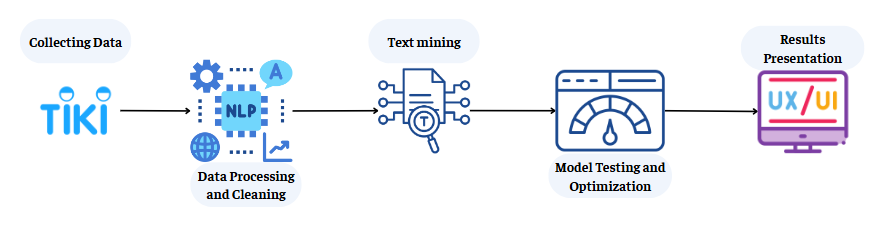
The research will be carried out through the following steps:

**Step 1: Data Collection**

Collect customer review data about books on the Tiki e-commerce platform, ensuring the data's relevance to the research topic.

**Step 2: Data Processing and Cleaning**

Study the data processing procedure, performing preprocessing steps such as removing duplicate reviews, handling missing values, and normalizing review text. Utilize natural language processing (NLP) techniques to clean and standardize the text data.



#### Figure 1.3. Implementation process

**Step 3: Text Mining**

Build and train a deep learning model, applying PhoBERT to analyze customer sentiment from the review data. Apply NLP algorithms to explore topics, perspectives, and customer satisfaction levels from the feedback in Vietnamese.

**Step 4: Model Testing and Optimization**

Conduct testing and fine-tuning of the model to improve the accuracy and efficiency of sentiment analysis. Evaluate the model's accuracy and reliability, ensuring the customer sentiment analysis model achieves high effectiveness.

**Step 5: Results Presentation**

Design and develop an interactive user interface that allows users to input reviews for classification and prediction by the model.

## 1.6. Report structure

**Chapter 1: Overview of the Project**

The first chapter provides an overview of the research project, including the rationale for selecting the topic, research objectives, scope of the research, and the subjects of the study. This chapter also presents the research methods, including data collection and the research implementation process.

**Chapter 2: Literature Review**

This chapter compiles and analyzes relevant research studies that have been previously published. The research team has searched for and analyzed articles and books with content that supports and provides recommendations for the project, based on relevant keywords.

**Chapter 3: Theoretical Framework**

This chapter presents the basic concepts and theories related to the research, including Deep Learning, BERT, Pho-BERT, NLP, and Sentiment Analysis. The methods and technologies used in the research are also explained in this chapter.

**Chapter 4: Proposed Model and Experiments**

Chapter 4 focuses on applying PhoBERT to the sentiment analysis task, including the processes of implementing the model and fine-tuning it for optimal performance. The chapter details the preprocessing of text data, feature engineering, and the application of various optimization techniques, such as hyperparameter tuning, to enhance model accuracy and efficiency.

**Chapter 5: Model Implementation**

This chapter focuses on deploying the model as an application, including the design of the user interface (UI) and the implementation process of the interface. The tools and techniques used to build the interface are also described in this chapter.

**Chapter 6: Conclusion**

The final chapter evaluates the performance of the model and presents the final conclusions of the research. This chapter also analyzes the new knowledge gained by the research team, assesses the contributions of the project, and identifies its limitations as well as potential directions for future development.

# CHAPTER 2. LITERATURE REVIEW

This chapter reviews previous research related to deep learning, text data analysis, PhoBERT, sentiment analysis in e-commerce. The chapter presents the models and methods that have been applied, and analyzes the results and limitations of these studies. Thereby, the chapter highlights the importance of applying sentiment analysis to better understand customer behavior, which forms the basis for our research on the Tiki platform.

## 2.1. Deep Learning

Referring to Deep Learning, the book “Dive into Deep Learning” [3] states that Machine learning is the study of algorithms that learn from experience. As a machine learning algorithm accumulates more experience, usually in the form of observational data or interactions with the environment, its performance improves. This is different from our pre-defined e-commerce platform, where business logic is applied uniformly, regardless of the amount of experience accumulated, until the developers themselves realize and decide that the software needs to be updated. The fundamentals of machine learning, especially deep learning, a powerful set of techniques that are driving innovation in fields as diverse as computer vision, natural language processing, healthcare, and genetics, show that the applications of Deep Learning and Machine Learning are extremely broad.

In the paper “Deep Learning” [4] by Ian Goodfellow (MIT University), it is said that deep learning is a method in machine learning that has relied heavily on knowledge of the human brain, statistics and applied mathematics over the past few decades. In recent years, deep learning has seen a significant increase in popularity and usefulness, mainly due to the development of more powerful computers, larger data and techniques for training deep neural networks. The coming years are full of challenges and opportunities to further improve deep learning and bring it to new areas.

For an example of an applied research on deep learning and its development, we can refer to the study “Towards Deep Learning Models Resistant to Adversarial Attacks” [5], which proposed a new method to enhance the resistance of deep learning models to adversarial attacks, using a robust optimization approach through the saddle point formulation. The authors demonstrate that this problem can be solved efficiently using the Projected Gradient Descent (PGD) method, and emphasize the importance of model capacity in creating robust neural networks. Experiments on the MNIST and CIFAR10 datasets show that this method can create models that are resistant to many types of adversarial attacks, opening up prospects for building secure neural networks in the future.

It can be seen that deep learning is becoming one of the most important innovative fields in computer science and artificial intelligence. With the ability to learn from data and improve performance through each experience, deep learning has gone beyond traditional methods, opening up many application opportunities in computer vision, natural language processing, and many other fields. The development of this technology is not only based on the increase in computing power and big data but also thanks to advanced techniques such as strong optimization and model improvement. Current research demonstrates the versatility and continued potential of deep learning in tackling complex challenges and improving model performance. Therefore, deep learning is not only an important technology in the present but also a foundation for future innovation and development, opening up many prospects for application in various fields and improving human life.

## 2.2. Customer sentiment analysis in e-commerce

The study “Customer Experience and satisfaction: Importance of customer reviews and Customer Value on buying preference” [6] concluded that restaurants should encourage customers to leave reviews on Zomato and respond to both positive and negative feedback. Zomato should consider implementing a feature that allows customers to filter reviews based on specific criteria, such as food quality or service. Restaurants should focus on improving their overall rating by addressing common complaints mentioned in negative reviews. From the recommendations, it can be seen that collecting and analyzing customer reviews is really necessary for the long-term development of the company.

Furthermore, the study “Determinants of Ecommerce customer satisfaction, trust and loyalty in Saudi Arabia” [7] investigated whether customer satisfaction and trust play a significant mediating role on Saudi Arabian consumers’ loyalty towards B2C e-commerce services. E-commerce customer satisfaction was found to have a significant influence on customer loyalty and to act as a partial mediator between each construct of user interface quality and information quality and customer loyalty.

In another study “The effect of customer feedback ratings on purchase decision in ecommerce” [8] that examined a multitude of factors that could potentially influence the amount of money customers are willing to spend, the study found that ratings and purchase decisions do not have an inverted U-shaped relationship but rather a polynomial relationship with purchase decisions. They found that a 1% increase in ratings would result in a customer being willing to pay 859.5 yen more. They also found that a 1% increase in rating leads to a 0.344 increase in utility.

In the increasingly competitive e-commerce market, analyzing customer reviews has become an extremely important factor. Customer reviews are not only a valuable source of information that helps businesses better understand their customers' experiences and needs, but also a powerful tool to build and maintain brand reputation. By analyzing customer feedback, businesses can detect problems in their products or services, thereby improving quality and enhancing customer satisfaction. Moreover, positive reviews can build trust and attract more potential customers, while promptly resolving negative reviews will help minimize the negative impact on the business's image. Therefore, analyzing customer reviews is not only a necessary task but also an important strategy to promote the sustainable development of businesses on e-commerce platforms.

## 2.3. Textual Data Analysis

The study “Sentiment Analysis of Top 5 E-commerce Platforms in Indonesia Using Text Mining and Natural Language Processing (NLP)” [9] applied textural data analysis and natural language processing to analyze user sentiments toward the top 5 e-commerce platforms in Indonesia based on the analysis of responses with 3 forms of 'positive', 'negative' and 'neutral'. In this paper, the important text mining techniques used include stopwords removal, term-document matrix generation, word frequency analysis and finally direct word cloud. Their results were given that 59.46% of users rated e-commerce platforms in Indonesia positively. The study also summarized the research methodology and how to apply textural data analysis to customer sentiment analysis.

On the other hand, the study “Study on Sentiment Analysis of Live Comments on E-commerce Based on Text Mining” [10] also uses text data analysis to analyze customer sentiments from live comments during live broadcasts on e-commerce platforms with the aim of mining information on users' emotional trends towards products, services and shopping experiences. This text data mining process is different from the previous study that builds an emotion dictionary and allocates emotional polarity values ​​for each word, combined with learning algorithms such as SVM, and Naive Bayes to train the model, the rest is in the same process as the previous study. The results of this study have contributed a deeper insight into users' emotions and attitudes and provided information for e-commerce enterprises to optimize business operations and improve emotion dictionaries.

In addition, the research paper “Model of opinion mining and online customer sentiment analysis in the food industry” [11] also proposed to produce a method of architectural opinion mining and online customer sentiment analysis in the food industry by collecting comments from the website Foody.vn. Using computational algorithms, the research analyzed text comments to mine architectural opinions and visualize the results, supporting business decision-making. The results showed that the accuracy of the method reached 90%, providing useful information to help stores and administrators better understand the advantages and disadvantages of products and services. The research also compared tactical studies to find the best model based on F-Score and generate direct reports. In the future, the research will expand the system of automatic updating and processing of big data, along with the development of mobile applications to support businesses and make more effective decisions.

## 2.4. Pho-Bert

The research topic “Sentiment Analysis Implementing BERT-based Pre-trained Language Model for Vietnamese” [12] focuses on analyzing sentiments from student feedback to improve the quality of education. The study uses the PhoBERT model, an optimized version of the BERT model for Vietnamese, to overcome the limitations of modern sentiment classification models that only focus on English. By applying fine-tuning techniques to extend the model for multi-class classification, the research method achieves excellent results on the UIT-VSFC dataset with an F1-score of 93.92% and an accuracy of 94.28%. This result not only supports the improvement of education in Vietnam but also opens up research opportunities for resource-poor languages ​​like Vietnamese.

Next, the paper "Vietnamese hate and offensive detection using PhoBERT-CNN and social media streaming data" [13] aims to develop an intelligent system to detect hateful and offensive content, thereby building a healthy and safe environment. The paper finds that current research in this field still has many shortcomings such as ineffective preprocessing techniques, disregard for data imbalance, modest model performance, and lack of practical applications. To overcome these problems, the authors propose some solutions. First, they introduce an effective preprocessing technique to clean comments from Vietnamese social networks. Then, they develop a new hate speech detection model, combining the pre-trained PhoBERT model and Text-CNN. EDA techniques are also applied to solve the data imbalance problem and improve the performance of classification models. Comparative experiments show that the PhoBERT-CNN model outperforms the state-of-the-art methods, achieving F1 scores of 67.46% and 98.45% on two benchmark datasets ViHSD and HSD-VLSP, respectively. Finally, they build an HSD live-streaming application to illustrate the practicality of the system. Overall, the paper focuses on developing the PhoBERT-CNN model for detecting hate speech in Vietnamese and demonstrates the effectiveness and practical application of the system through experiments and real-world applications.

Finally, the paper "Combining PhoBERT and SentiWordNet for Vietnamese Sentiment Analysis" [14] focuses on the task of sentiment analysis, one of the most important tasks in Natural Language Processing (NLP), where machine learning models are trained to classify texts according to the positivity of the sentiment. Many models have been proposed to solve this task, among which the pre-trained PhoBERT models are considered the most advanced for Vietnamese. PhoBERT's pre-training method is based on RoBERTa, which optimizes BERT's pretraining method to achieve stronger performance. In this paper, the authors introduce a new method that combines PhoBERT and SentiWordNet to analyze sentiments from Vietnamese reviews. The proposed sentiment analysis model uses PhoBERT, a robust optimization for Vietnamese from the prominent BERT model, and SentiWordNet, a lexical resource specifically designed to support sentiment classification applications. Experimental results on the VLSP 2016 and AIVIVN 2019 datasets show that their sentiment analysis system achieves good performance compared to other models.

# CHAPTER 3. THEORETICAL BACKGROUND

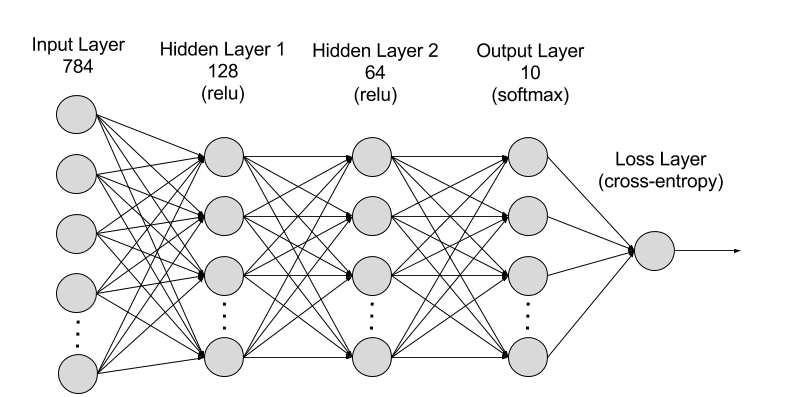
Chapter 3 provides the theoretical foundation for the research, including concepts related to deep learning, particularly models like BERT and PhoBERT, as well as natural language processing. It discusses the effectiveness of these models, laying the groundwork for the subsequent research.

## 3.1. Deep Learning

Deep learning is a branch of artificial intelligence (AI) that mimics the way the human brain works to process and analyze data. Deep learning models are capable of recognizing and classifying complex patterns in images, text, audio, and other forms of data, generating insights and accurate predictions. This technology can automate complex tasks such as describing images or converting audio to text.

Deep learning models play a key role in the development of artificial intelligence (AI) by giving computers the ability to learn and think like humans. This technology is widely used in everyday products and services such as digital assistants, voice control, fraud detection, and facial recognition. Deep learning also underpins advanced technologies such as self-driving cars and virtual reality.

Deep learning has many applications in fields such as automotive, aerospace, healthcare, and manufacturing. For example, self-driving cars use deep learning to recognize traffic signs and pedestrians; defense systems use it to analyze satellite images; medicine uses it to detect cancer cells in diagnostic imaging. Other applications include speech recognition, natural language processing, and personalized recommendation engines.

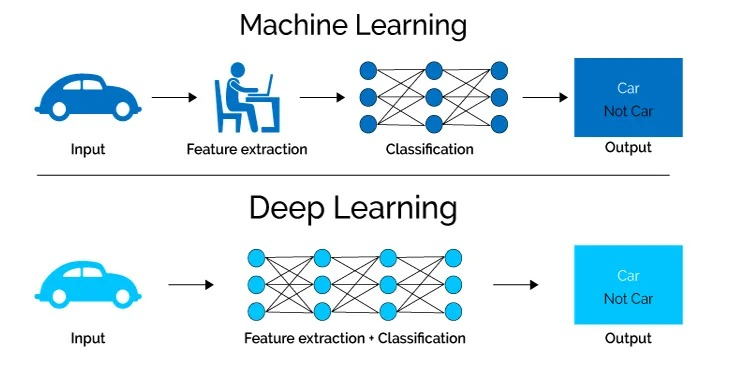


#### Figure 3.1. Deep neural network model [15]

Deep learning models are based on artificial neural networks, which simulate the structure of the human brain. This neural network consists of input layers, hidden layers, and output layers. Hidden layers process information at different levels, allowing deep learning networks to analyze and make predictions from complex data.

A deep learning neural network consists of an input layer to receive data, a hidden layer to process data, and an output layer to output results. Hidden layers can have hundreds of layers, each layer processes a different aspect of the data, helping deep learning networks analyze problems from multiple perspectives and produce more accurate results.

Deep learning is a part of machine learning, born to improve the efficiency of traditional machine learning techniques. Traditional machine learning methods require a lot of effort in labeling and processing data. In contrast, deep learning models have the ability to learn by themselves and improve over time without needing too much labeled data.



#### Figure 3.2. How Machine Learning and Deep Learning work [16]

Compared to machine learning, deep learning models can efficiently process unstructured data, discover hidden relationships and patterns in data, and learn unsupervised. This allows deep learning models to analyze data more deeply, make more accurate predictions, and automatically improve over time without much manual intervention.

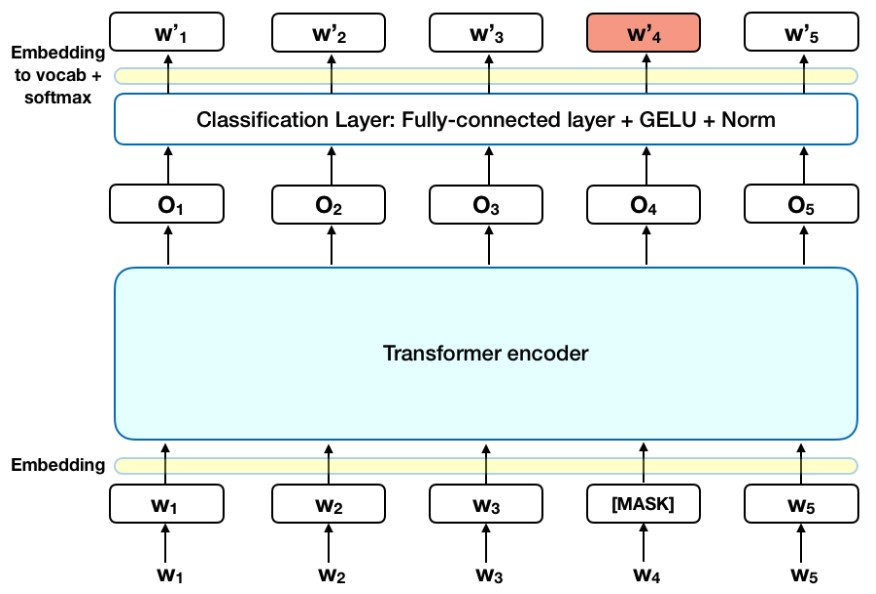
Some of the challenges of deep learning include the need for large amounts of high-quality data and large processing power. Input data needs to be cleaned and processed before it can be used to train deep learning models. In addition, deep learning algorithms require powerful infrastructure to operate effectively, otherwise the processing can be very time-consuming.

## 3.2. Bert

BERT (Bidirectional Encoder Representations from Transformers) is a language model developed by Google AI. BERT marks a major breakthrough in the field of Machine Learning thanks to its ability to be applied to many natural language processing (NLP) problems such as question answering and natural language inference, with very high efficiency.

One of the biggest challenges of NLP is handling heterogeneous data. There is a huge amount of data on the Internet, but they are often used for separate purposes. This leads to us having only a small amount of data suitable for each specific problem, while Deep Learning models need a large amount of data to achieve good results. Transfer Learning, a new technique, helps to overcome this problem by building general models from huge data on the Internet and fine-tuning them for each specific problem. BERT is an excellent representative of Transfer Learning in NLP, bringing high efficiency and is completely free.

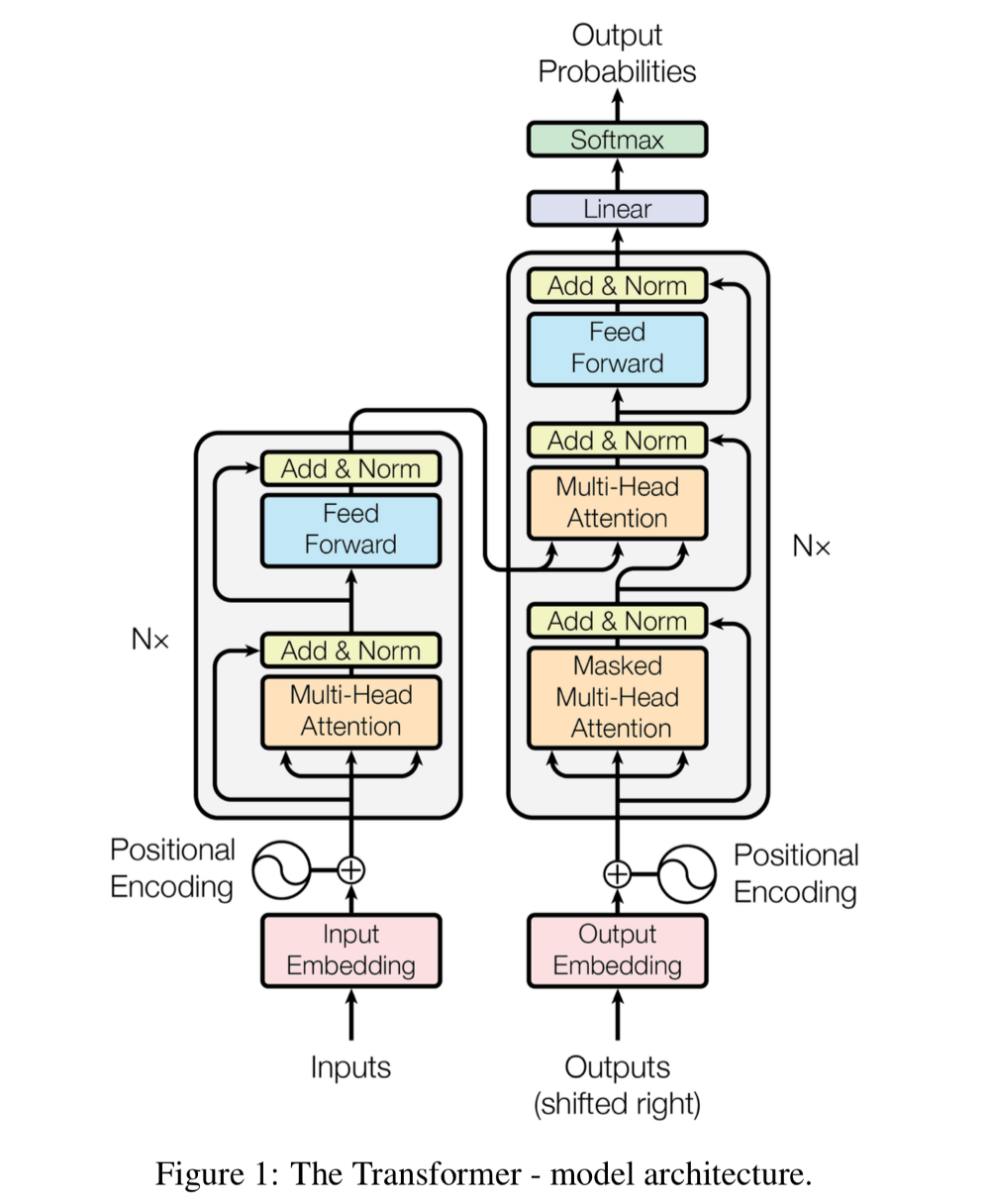
BERT works on the Transformer, a model that uses attention to learn the correlation between words in a text. The Transformer consists of two main parts: Encoder and Decoder, in which BERT only uses the Encoder part. Unlike directional models that only read data in one direction (left to right or vice versa), BERT reads the entire data at once, helping the model learn the context of words better by using information from both sides.



#### Figure 3.3. BERT model with Masked LM [17]

The way the Encoder works in BERT starts with the Encoder's Input, which is a sequence of tokens represented as a sequence of vectors. The output of the model is a sequence of vectors with the same size as the input. During training, BERT uses two main strategies to overcome the limitations of traditional directional models: Masked LM (MLM) and Next Sentence Prediction (NSP).

With Masked LM (MLM), 15% of the words in the sequence are replaced with [MASK] tokens before being fed into BERT. The model predicts the replaced word based on the context of the surrounding words. This process involves adding a classification layer, multiplying the output vectors with the embedding matrix, and calculating the probability of each word in the vocabulary using softmax. BERT's error function only focuses on evaluating [MASK]-tagged words, which helps the model understand context better, although it converges more slowly.



#### Figure 3.4. Transformer model architecture [18]

As for Next Sentence Prediction (NSP), in this strategy, the model uses a pair of sentences as input data and predicts whether the second sentence is the next sentence of the first sentence. The training data consists of 50% consecutive sentence pairs and 50% random sentence pairs. The processing steps include inserting [CLS] and [SEP] tokens, labeling the tokens in each sentence, and adding an embedding vector representing the position of the token in the sentence. Then, the entire input sentence is fed into the Transformer to calculate the IsNextSequence probability using softmax.

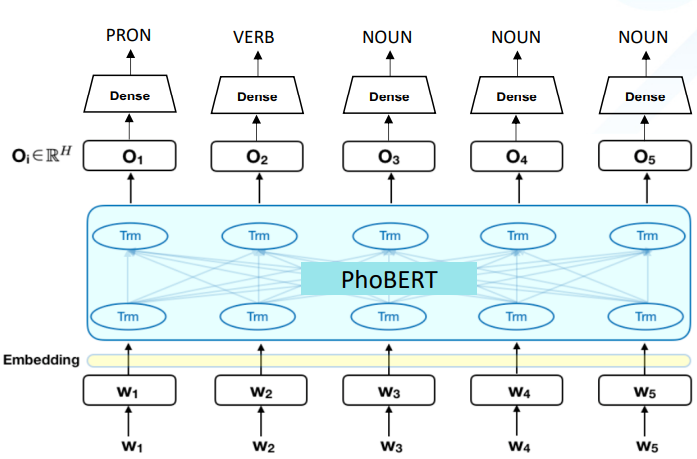
A special point is that BERT can be fine-tuned for many different problems. For the Classification problem, we add a classification layer with the input being the output of the Transformer for the [CLS] token. For the Question Answering problem, the model receives a text segment and a question as input, and is trained to label the answer in the text segment. In the Named Entity Recognition (NER) problem, the model is trained to predict the label for each token, such as person name, organization name, place name, etc.

## 3.3. Pho-Bert

PhoBERT is an advanced language model specifically designed for Vietnamese. Developed by VinAI Research, PhoBERT is the most optimized version of the BERT model specifically for Vietnamese, helping to solve many natural language processing (NLP) problems such as text classification, sentiment analysis, and entity labeling. The name "PhoBERT" is inspired by the famous Vietnamese dish Pho, reflecting the uniqueness and localization of the model.

Vietnamese has a complex grammatical and lexical structure, leading to the application of general language models not being very effective. Before PhoBERT, researchers had to use non-specific models, leading to inaccurate results. PhoBERT overcomes this by being trained on a large amount of Vietnamese data, helping the model gain a deeper understanding of the context and grammatical structure of Vietnamese. This significantly improves the efficiency in NLP problems related to Vietnamese.

PhoBERT is based on the Transformer architecture like BERT, using the attention mechanism to learn the relationship between words in a sentence. There are two versions of PhoBERT: PhoBERT-base and PhoBERT-large, similar to BERT, with different numbers of parameters to suit different requirements and computational resources. PhoBERT uses a pre-training method based on RoBERTa, optimizing the training process of BERT to achieve higher performance.



#### Figure 3.5. PhoBERT model [19]

PhoBERT uses the Masked Language Model (MLM) strategy in the training process, similar to BERT. In MLM, a part of the word in the sentence is replaced by the token [MASK], and the model has to predict the masked word based on the surrounding context. This helps the model learn the context and relationship between words better.

PhoBERT has been widely applied in many Vietnamese NLP problems and has achieved remarkable results. PhoBERT has set a new standard for Vietnamese NLP tasks such as part-of-speech tagging, dependency parsing, named-entity recognition, and natural language inference. PhoBERT outperforms previous monolingual and multilingual methods, achieving the best performance on Vietnamese NLP tasks. This demonstrates that a language model trained specifically for a specific language can produce better results than non-specialized models. PhoBERT is not only a powerful tool for researchers and developers in the field of NLP, but also a testament to the progress in developing language models for less-noticed languages.

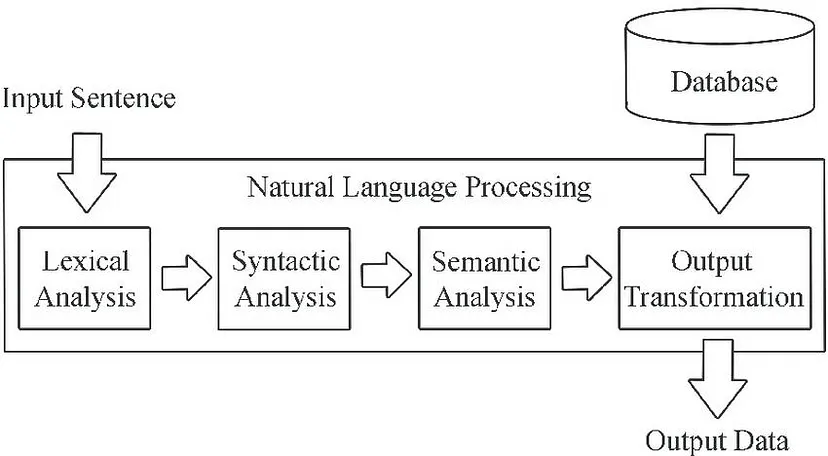
## 3.4. Natural Language Processing (NLP)

Natural Language Processing (NLP) is a field of computer science and artificial intelligence that focuses on giving computers the ability to interpret, understand, and interact with natural human language. NLP technology enables the processing and analysis of large volumes of voice and text data from various communication channels such as emails, text messages, social media feeds, video files, audio files, and more. Using NLP software, organizations can automatically process and analyze this data, identify the intent or sentiment in messages, and respond effectively.

NLP plays a vital role in effectively analyzing a full range of text and voice data. It is capable of handling dialectal differences, slang, and grammatical irregularities that are common in everyday conversations. Companies use NLP technology for a variety of automated tasks such as processing, analyzing, and storing large documents, analyzing customer feedback or call center recordings, running chatbots for automated customer service, and classifying and extracting text.

NLP can be integrated into customer-facing applications to communicate more effectively. For example, chatbots use NLP to analyze and classify customer queries, automatically answer frequently asked questions, and route complex queries to customer support. This reduces costs, saves employees time, and improves customer satisfaction.

Businesses use NLP to simplify, automate, and streamline operations. For example, in the insurance, legal, and healthcare sectors, NLP technology helps redact personally identifiable information and protect sensitive data. In customer interactions, chatbots and voice bots use NLP to communicate more human-like, improving customer service. Marketers use NLP to analyze customer sentiment towards a company's products or services, thereby coming up with appropriate business strategies.



#### Figure 3.6. Steps in the NLP process [20]

NLP combines computational linguistics, machine learning, and deep learning models to process human language. Computational linguistics studies and builds human language models using computer tools and software. Machine learning trains computers with sample data to improve efficiency. Deep learning, a branch of machine learning, focuses on teaching computers to learn and think like humans through neural networks.

The NLP implementation process begins with the collection and preparation of text or speech data from various sources. NLP software uses pre-processing techniques such as tokenization, word shortening, infinitive recovery, and stop word removal to prepare the data for various applications. Researchers then use the pre-processed data and machine learning to train NLP models. Finally, these models are deployed or integrated into existing production environments to receive input data and make predictions for specific tasks.

NLP tasks include part-of-speech labeling, word sense ambiguity resolution, speech recognition, machine translation, entity recognition, and sentiment analysis. For example, part-of-speech labeling helps computers understand how words form semantic relationships with each other in sentences. Entity recognition identifies unique names for people, places, events, companies, and more. Sentiment analysis helps interpret the emotions conveyed through text data.

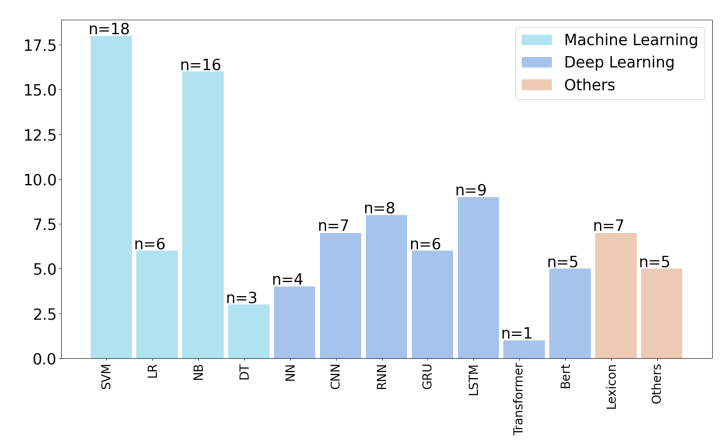
NLP can be implemented in a variety of ways, including supervised and unsupervised methods. Supervised methods train software with labeled input and output data sets, while unsupervised methods use statistical language models to predict patterns that occur when given unlabeled input data. Natural language understanding (NLU) and natural language generation (NLG) are two sub-branches of NLP that focus on analyzing the meaning behind sentences and generating human-like conversational text.

## 3.5. Sentiment Analysis

Sentiment Analysis is an important field in natural language processing (NLP) aimed at identifying and classifying emotions expressed in text. With the rapid development of online platforms, understanding user sentiment has become extremely important for organizations to make informed decisions.

Sentiment analysis can be applied in various fields such as business, finance, and in this study, retail. Understanding customer sentiment about products and services helps improve customer satisfaction and enhance brand reputation. For example, in the retail sector, sentiment analysis from articles, posts, or customer comments on social media or e-commerce sites can help businesses assess consumer perceptions of products and develop better strategies.

Sentiment analysis is based on various techniques such as machine learning, deep learning, and ensemble learning. These techniques include data preprocessing, feature extraction, and classification. Deep learning models like LSTM and CNN have been widely applied and achieved high performance in sentiment analysis.



#### Figure 3.7. Commonly used machine learning and deep learning models [9]

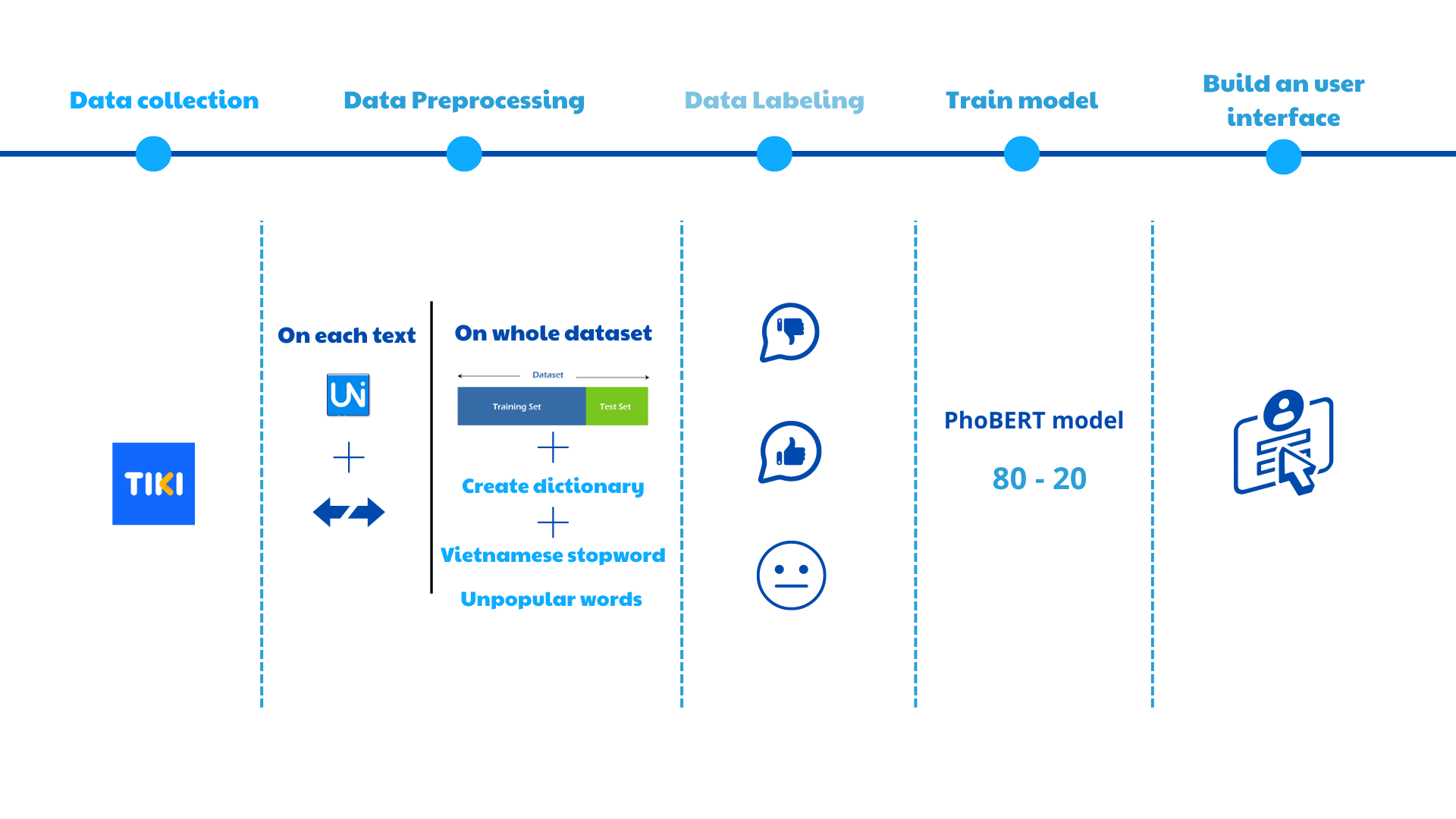
One of the major challenges of sentiment analysis is the complexity of natural language and the diversity of emotional expressions. Studies have shown that using language-specific models can significantly improve the effectiveness of sentiment analysis [21]. Therefore, to gain the most insightful view of sentiment classification through Vietnamese comments, this study leverages PhoBERT [19] - a pre-trained Vietnamese language model by VinAI to optimize customer sentiment analysis.

Sentiment analysis is not only a powerful tool for researchers and developers in the NLP field but also a testament to the progress in developing language models for less commonly studied languages worldwide.

# CHAPTER 4. PROPOSED MODEL AND EXPERIMENT

This chapter describes the process of implementing the PhoBERT model for sentiment analysis on Tiki's comment data. It begins with the steps for data preparation, including preprocessing and dataset splitting. The chapter then explains how to configure the model and the parameters used during training. Finally, it presents the performance evaluation results of the model, including accuracy, sensitivity, and other metrics, to assess the effectiveness of the sentiment analysis.

## 4.1. Problem and proposed model



#### Figure 4.1. Proposed model (proposed by the research team)

In the context of the growing development of e-commerce and social media platforms, classifying sentiments from customer comments has become extremely important. To do this effectively, we use the PhoBERT model, a BERT-based language model optimized for Vietnamese. Below are the data preprocessing steps to train this model to classify sentiments from customer comments.

First, the team will collect data from data sources collected from many different sources such as social networks, forums, and product review websites. Each comment must be accompanied by a corresponding sentiment label, such as positive, negative, or neutral. This helps ensure that the model is trained on diverse and rich data.

After successful collection, the team proceeds to preprocess the data with a series of text data preprocessing operations ranging from word level to the entire dataset. For each text level, the research team performs processing operations such as Unicode standardization, vowel standardization, punctuation, etc. For preprocessing the entire dataset, the research team re-divides the training set and the test set and performs basic text data processing operations such as removing stop words, special characters, uncommon words, etc.

Next, each comment will be labeled and formatted according to user ratings and reviews. This makes loading data into the PhoBERT model easy and convenient.

Then, the most important step of the project will be training the sentiment classification model using the PhoBERT deep learning method, and the training process for the model will be carried out. In this step, the train, test, and tokenizer data will be separated to convert the text into tokens that the model can understand. Next is to train the emotion classification model based on the training dataset, and evaluate the model performance with the test dataset.

In the final step to complete the project, the team will build an interactive interface with the user to apply the emotion classification model to the design and build an interactive interface with the user.

## 4.2. Data collection method

The data set used in the research includes a data set of books posted for sale on the Tiki e-commerce platform and a data set of comments evaluating the quality of book products that they used to consume. The research team inherited data from a publicly posted study on the Kaggle.com platform titled Tiki Books Dataset [22]. The dataset includes small data files containing information about 2024 bestselling books and customer reviews on the Tiki e-commerce platform.

The original data includes 3 files as follows:

##### Table 4.1. Tiki Books Dataset description

| **File name** | **Column name** | **Description** |
| --- | --- | --- |
| book\_data | product\_id | ID of the product in the Tiki database (unique) |
| title. title | Name of the book, maybe contain republish time |
| authors | Same with its name |
| original\_price | Price at the first time |
| current\_price | Price at present if having a discount |
| quantity. quantity | Total number of books sold of all time |
| category. category | Kind of books |
| n\_review | Number of reviews |
| avg\_rating | Average rating (max 5.0) |
| pages. pages | Total pages of each book |
| manufacturer. manufacturer | Name of publisher |
| cover\_link | Link to the website to buy the book on Tiki |
| book\_id | id | ID of all products when I crawl. You can ignore |
| comments. comments | product\_id | Same with book\_data file |
| comment\_id | Each comment has individual id |
| title. title | Keyword of comments |
| thank\_count | Number of likes of other people |
| customer\_id | Each customer has individual id |
| rating. rating | Average rating of the comments |
| content. content | Same with it's name |

## 4.3. Data preprocessing

### 4.3.1. Preprocessing of assessment data

* **Convert the acronym back to the original**

First, the research team defined acronyms or heart sounds and converted them to their original full state such as no - no, ok - okay,...

Some abbreviations and slang words were analyzed and defined by the research team:

##### Table 4.2. Definition of abbreviations and slang words

| **Abbreviations** | **From the root** |
| --- | --- |
| no | Are not |
| k | Are not |
| Are not | Are not |
| no yes | Not available |
| cough yes | Not available |
| cx | also |
| ok | Okay |
| ntn | How |
| r | Already |
| vs | with |
| bt | Normal |
| dc | Okay |
| ok | Okay |
| tl | reply |
| thk | prefer |

After defining and building the function, the initial comment data was preprocessed in terms of acronyms and slang words appearing in the sentence. By building a function that handles acronyms and slang words:

* Use a for loop to iterate through pairs of acronyms and their full forms in the dictionary
* Use the re.sub function from python's re (Regular Expression) library to search and retrieve “abbr” into its “full\_form” within the dictionary
* Additionally, construct a regular expression to ensure that chit replaces exact abbreviations (separated by word boundaries “\b” that are not part of another word. For example, the function will only is triggered when “ASAP” stands alone, but if “ASAP” appears in other words like “ASAPly”, the conversion function will not work.
* **Encode punctuation and vowels**

To perform vowel encoding in Vietnamese documents, the research team manually built a library of vowel tables and punctuation character tables in Vietnamese.

##### Table 4.3. Definition of vowels

| **Vowels have accents** | **Definition of vowels** |
| --- | --- |
| 'a', 'ah', 'ah', 'ah', 'ã', 'ah' | 'a' |
| 'ã', 'ã', 'ã', 'ã', 'ã', 'ã' | 'aw' |
| 'â', 'à', 'ã', 'ã', 'ã', 'ã' | 'aa' |
| 'e', 'è', 'é', 'ê', 'ê', 'ê' | 'e' |
| 'eh', 'eh', 'eh', 'eh', 'eh', 'eh' | 'ee' |
| 'i', 'ì', 'í', 'ì', 'ì', 'poo' | 'i' |
| 'o', 'ò', 'o', 'o', 'õ', 'ọ' | 'o' |
| 'oh', 'oh', 'oh', 'oh', 'oh', 'oh' | 'oo' |
| 'uh', 'uh', 'uh', 'here', 'uh', 'burp' | 'ow' |
| 'u', 'ù', 'ú', 'nu', 'ũ', 'ô' | 'u' |
| 'uh', 'uh', 'ugh', 'uh', 'uh', 'uh' | 'uw' |
| 'y', 'uh', 'y', 'y', 'uh', 'y' | 'y' |

##### Table 4.4. Name of Punctuation characters

| **Punctuation name** | **Punctuation characters** |
| --- | --- |
| ' ' | Horizontal |
| 'f' | Apostrophe |
| 'S' | Acute |
| 'r' | Question mark |
| 'x' | Tilde |
| 'j' | Underdot |

* For punctuation encoding

Here, the research team uses the normalize function in python with the "unicodedata" module to normalize Unicode, specifically normalizing text into "Normalization Form Composed" - NFC. That is, each punctuation mark will be displayed using the Unicode code of the marks when combined during the NFC training process.

* For vowel encoding

First, the research team checked the validity of Vietnamese words in documents by considering the position of vowels in words. A valid Vietnamese word is a word with vowels standing next to each other according to Vietnamese grammar rules. Here, there is a special case that the research team proposes needs to be standardized related to vowels: special cases with words with "qu" or "gi".

Handling special cases with only 1 vowel or "qu", "gi"

First, perform a test and identify words with “qu” or “gi”. After determining:

* Words with many vowels: Put appropriate punctuation on those words by defining the position of that vowel in the word and placing punctuation in the appropriate position if any.
* Words with 2 vowels: Punctate the appropriate vowel (usually the last vowel if it is not the last vowel of the word).
* Words with 3 or more vowels: Put punctuation on the 2nd vowel

Finally, combine the standardized characters into complete words, Handling punctuation, Combine operations that convert all words to normal and split the sentence into a list of words based on spaces. Each word will be processed separately.

The research team separates words from surrounding accents:

* Part 1: Start with non-letter characters (punctuation) at the beginning of the word.
* Part 2: Main part of the word (alphabetic characters).
* Part 3: Non-letter characters (punctuation marks) at the end of words.

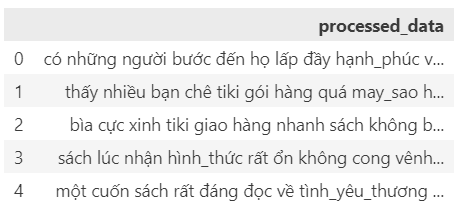
Perform a check on the split word if the character in the word being considered has an accent, save this accent in the variable dau\_cau and replace it with an unaccented vowel built in advance by the research team.

* **Convert into regular words and separate words**

After completing the punctuation encoding for vowels in the text, the research team converted all characters to non-uppercase lowercase form and separated words in the text.

* **Standardize sentences**

For words containing unwanted accented vowel characters such as accented vowels in the previously built vowel point group, remove unwanted characters and replace extra spaces. Convert them to a single space.

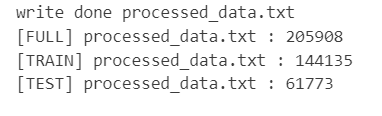


#### Figure 4.2. Data after preprocessing

### 4.3.2. Preprocess data across the entire dataset

* Standardize and export data into train and test sets

In this step, initialize the list of labels (columns) in the entire data set. Each column name will represent a label (category).



#### Figure 4.3. Number of data rows in each set

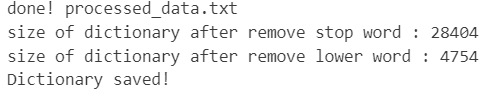
With some review data lines that are duplicated in their IDs, the research team extracted and removed the duplicated review data lines and kept only one row. In addition, the appearance of empty review data lines makes the data unbalanced, which can easily affect the performance of model training. The research team also removed review lines with empty “content” columns.

With that list of categories, the research team transfers the labels to the appropriate data set (training set or test set) and then creates two files for training and testing with corresponding paths.

* Create dictionary from text data

From the previously divided training set, the research team created a dictionary of words in the text file, with each word being a key and the value of the key being the number of occurrences of that word. The research team obtained results about the size of the dictionary, i.e. the number of different words in the dictionary is 28743.

After synthesizing and creating the dictionary, the research team performs dictionary processing by removing stop words and words that are not common or contain numbers, then saves the final dictionary to a file.



#### Figure 4.4. Dictionary after standardization

* Remove stop words

Use the Vietnamese stop word file publicly provided on many *vietnamese-stopwords* [23] platforms to filter out words that are not in this list and create a new dictionary. And the size of the dictionary after this step reduces to 28404.

* Remove less common words and words containing numbers from the dictionary

Here, the research team creates a new dictionary that only contains words with a number of occurrences greater than or equal to 30 and does not contain numeric characters. And the size of the dictionary after this step reduces to 4754.

## 4.4. Topic Labeling

Before performing topic labeling, the text data is prepared through loading and preprocessing. First, a label column ("Label") is created based on the star rating column ("Rating"). Specifically, 1-2 star reviews are labeled "negative", 3-star reviews are labeled "neutral", and 4-5 star reviews are labeled "positive". Next, only reviews with titles that fall into valid categories such as "Extremely satisfied", "Satisfied", "Normal", "Very dissatisfied", and "Dissatisfied" are retained, while other titles are discarded.

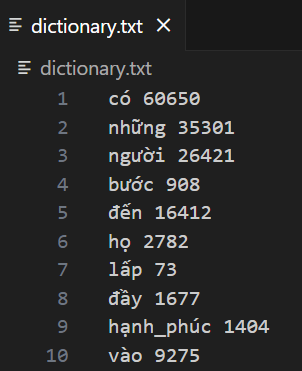
To ensure the accuracy of the data, a sentiment filter is applied to remove cases where reviews have positive titles but a default rating of low because the user forgot to change it. Specifically, this filter only retains positive reviews when the title is "Extremely Satisfied" or "Satisfied" and the label is "Positive"; neutral reviews when the title is "Neutral" and the label is "Neutral"; and negative reviews when the title is "Extremely Dissatisfied" or "Dissatisfied" and the label is "Negative". Then, records with duplicate "comment id" are removed to ensure that each review is unique. Finally, the processed data is saved as a file 'last\_dataset.csv' for use in the next steps of the analysis and model building process. The dataset 'last\_dataset.csv' contains product reviews with relevant information such as title, number of thanks, star rating, and review content.

After loading the data, the column 'processed\_data' is renamed to 'train' to serve as input data for the processing. Sentiment labels (‘labels’) such as “positive”, “neutral”, and “negative” are also converted into numeric values ​​2, 1, and 0 respectively to prepare for the machine learning model.

##### Table 4.5. Topic label definitions

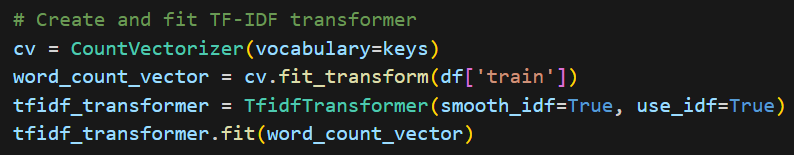
| **Topic Label** | **Topic Name** | **Brief Description** |
| --- | --- | --- |
| 0 | Negative | Negative emotions |
| 1 | Neutral | Neutral emotions |
| 2 | Positive | Positive emotions |

The topic labeling process uses the TF-IDF (Term Frequency-Inverse Document Frequency) method to represent documents as a vector of numbers. First, a vocabulary is created from the list of words in the file ‘dictionary.txt’.



#### Figure 4.5. Dictionary of dataset

This dictionary is then used to count the frequency of words in each document using a ‘CountVectorizer’. The resulting word count matrix is ​​converted into a TF-IDF matrix, which represents the importance of each word in the document relative to the entire dataset.



#### Figure 4.6. Create and fit TF-IDF transformer

Each line of text after processing will have its TF-IDF calculated by calling the function ‘computing\_tfidf’, which uses the previously trained ‘CountVectorizer’ and ‘TfidfTransformer’ models to convert the text into a TF-IDF vector.



#### Figure 4.7. Some lines of data after TF-IDF calculation

Each cell in the tfidf column is a Python dictionary with keywords being words in the document and values ​​being the TF-IDF scores of that word. The components such as “Keywords” are words in the document that are entered into the dictionary as keys, “TF-IDF Score” are values ​​corresponding to each word that indicate the importance of that word in the document. This score is calculated based on a combination of the word’s frequency of occurrence (Term Frequency - TF) and the word’s rarity in the entire dataset (Inverse Document Frequency - IDF).

How to Calculate TF-IDF Score Specifically:

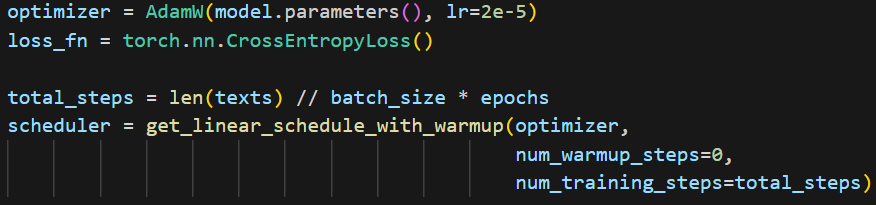
1. Term Frequency (TF): Measures the frequency of a word in a document. TF = (Number of times the word appears in the document) / (Total number of words in the document).
2. Inverse Document Frequency (IDF): Measures the rarity of a word in the entire dataset. IDF = log((Total number of documents) / (Number of documents containing the word)).
3. TF-IDF Score: Is the product of TF and IDF. The formula is:

TF-IDF = TF × IDF

## 4.5. Model Training

To build the sentiment classification model, the team used a combination of the PhoBERT language model and TF-IDF features. PhoBERT is a language model pre-trained on Vietnamese data, which helps the model understand the context of words in a sentence accurately.

The sentiment classification model is built based on the ‘SentimentClassifier’ layer. In this layer, first is the PhoBERT model, then a dropout layer to minimize overfitting, and finally two fully connected layers to combine PhoBERT features with TF-IDF features, before making the final prediction.



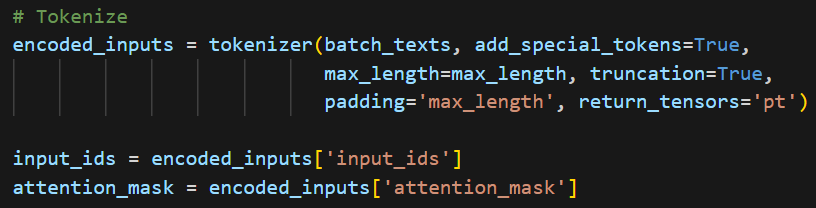
#### Figure 4.8. Initialize training components and learning schedule

The model is trained with a dataset that has been split into a training set and a testing set. The text features are tokenized and converted into a format suitable for the PhoBERT model. At the same time, TF-IDF features are also prepared to combine with PhoBERT language features. The training process uses ‘AdamW’ as the optimizer, with the loss function ‘CrossEntropyLoss’. To adjust the learning rate, ‘get\_linear\_schedule\_with\_warmup’ is used.

* **`optimizer`:** Create an AdamW optimizer object with a learning rate of 2e-5. This is the optimization algorithm used to update the model's weights.
* **`loss\_fn`:** Create a CrossEntropyLoss loss function, often used for multi-class classification problems.
* **`total\_steps`:** Calculate the total number of training steps based on the number of documents, batch size, and number of epochs.
* **`scheduler`:** Create a get\_linear\_schedule\_with\_warmup object to adjust the learning rate with each training step.

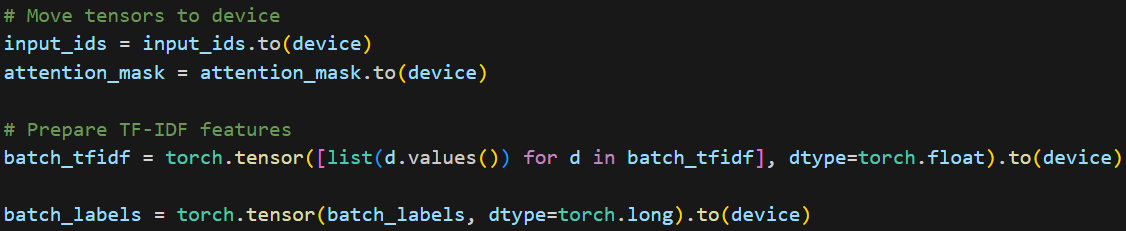
The model is trained over multiple epochs, and the model performance is evaluated through the accuracy and loss after each epoch. In each epoch, the model performs the following training steps:

* **Text tokenization**



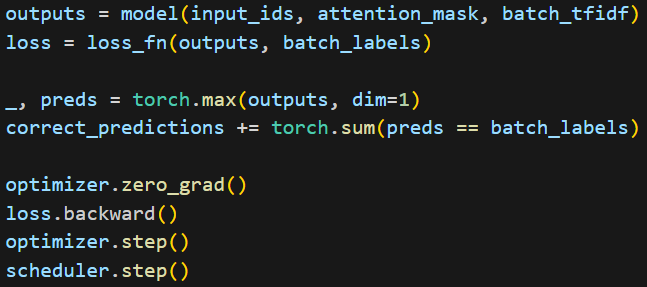
#### Figure 4.9. Text Tokenization Processing

* **`encoded\_inputs`**: Tokenize text in batch, returning word indices and attention mask.
* **`input\_ids`**, **`attention\_mask`**: The input tensors and attention masks are passed to the device (CPU or GPU).
* **Convert TF-IDF to tensor**



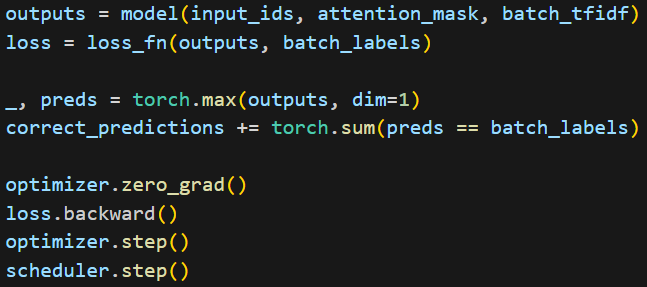
#### Figure 4.10. TF-IDF feature conversion process to tensor

* **`batch\_tfidf`**: Convert TF-IDF feature into tensor and feed into device.
* **`batch\_labels`**: Convert the label to a tensor and feed it to the device.
* **Prediction and loss calculation**



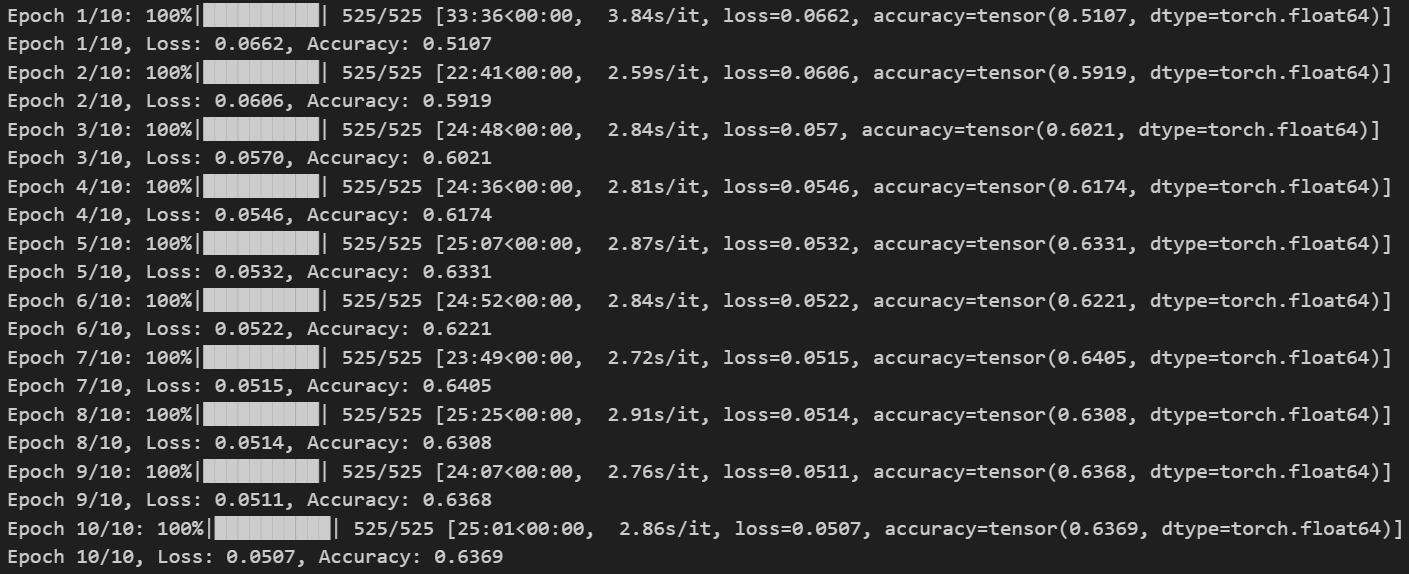
#### Figure 4.11. Prediction and loss calculation

* **`outputs`**: Calculate predictions from the model with input `input\_ids`, `attention\_mask`, và batch\_tfidf.
* **`loss`**: Calculate the loss between the prediction and the actual label.
* **Update model weights**



#### Figure 4.12. Calculate correct predictions and update weights

* **`preds`**: Get the model's prediction (the class with the highest probability).
* **`correct\_predictions`**: Cumulative number of correct predictions.
* **`optimizer.zero\_grad()`**: Set the gradient of the weights to 0.
* **`loss.backward()`**: Calculate the gradient of the loss.
* **`optimizer.step()`**: Update the model weights.
* **`scheduler.step()`**: Update the learning schedule.

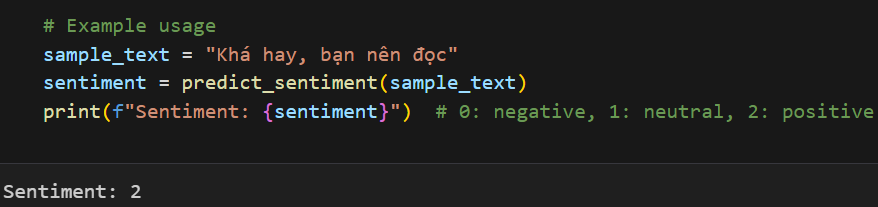


#### Figure 4.13. Result of training model

After training, the model is evaluated on the test dataset to determine its performance. The model's accuracy improves with each epoch, indicating that the model is gradually learning important features of the data. The model that achieves the best results is saved for future prediction use.

The training results showed that the model achieved an accuracy of 51% to 64% over epochs, indicating that the model improved over time.

Finally, the team used a function predict\_sentiment, which was built to test the model with new text passages. This function tokenizes the text, calculates the TF-IDF feature, and then uses the trained model to predict the sentiment of that text.



#### Figure 4.14. Test and predict model with example text

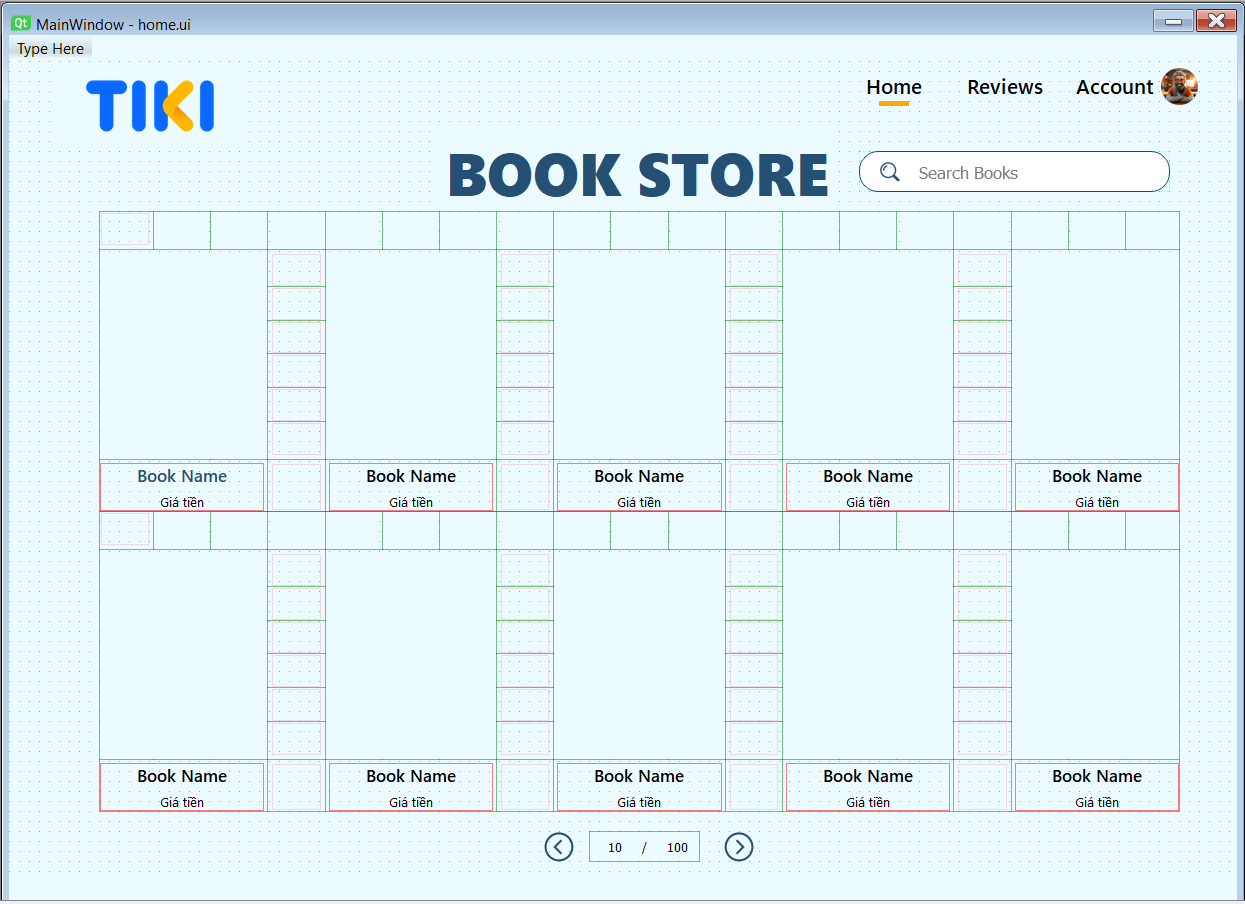
# CHAPTER 5. MODEL APPLICATION

Chapter 5 describes the implementation of the sentiment analysis application following the training of the model. The application features a user interface and presents the results of the testing, along with user feedback, to evaluate the effectiveness of both the model and the interface.

## 5.1. User Interface

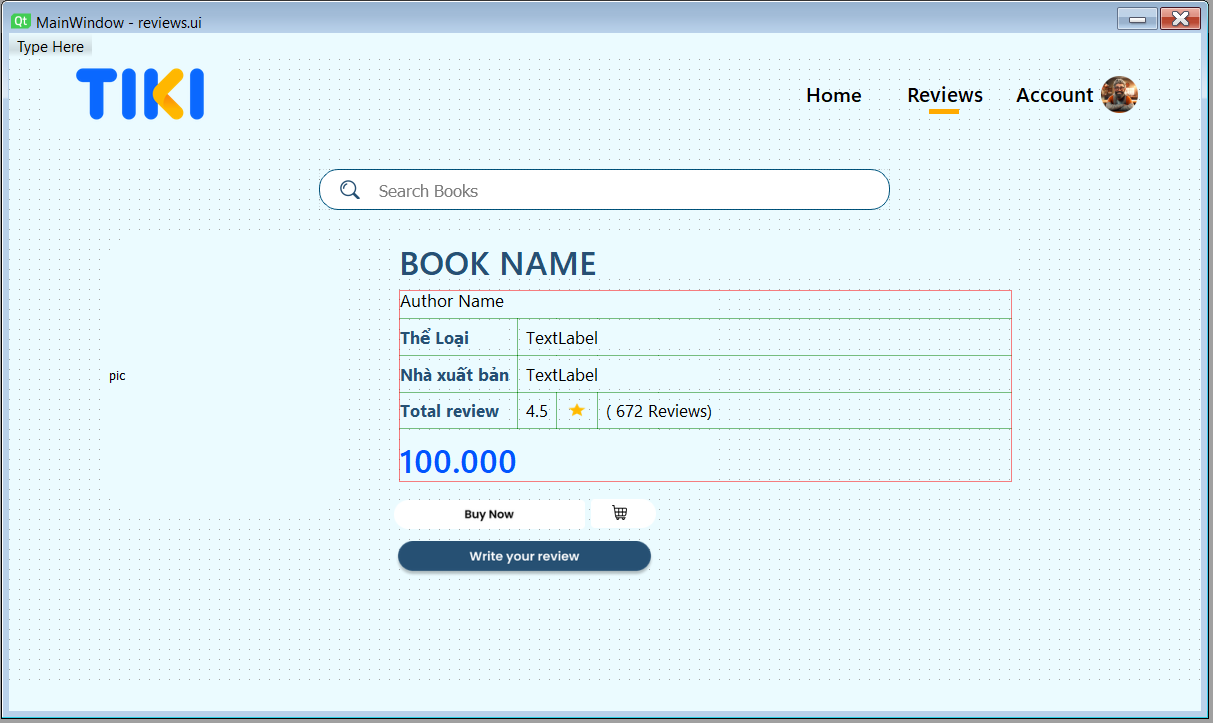
After training the model, the research team designed a visual application to present the application in an easy-to-understand way. For the User Interface (UI) design, the team used Qt Designer software to design the application.

For the home page of the application, the team designed a bookstore that displays information about the book title, price, and cover image. In addition, a book search section is also displayed so that customers can search for books that suit their interests.

****

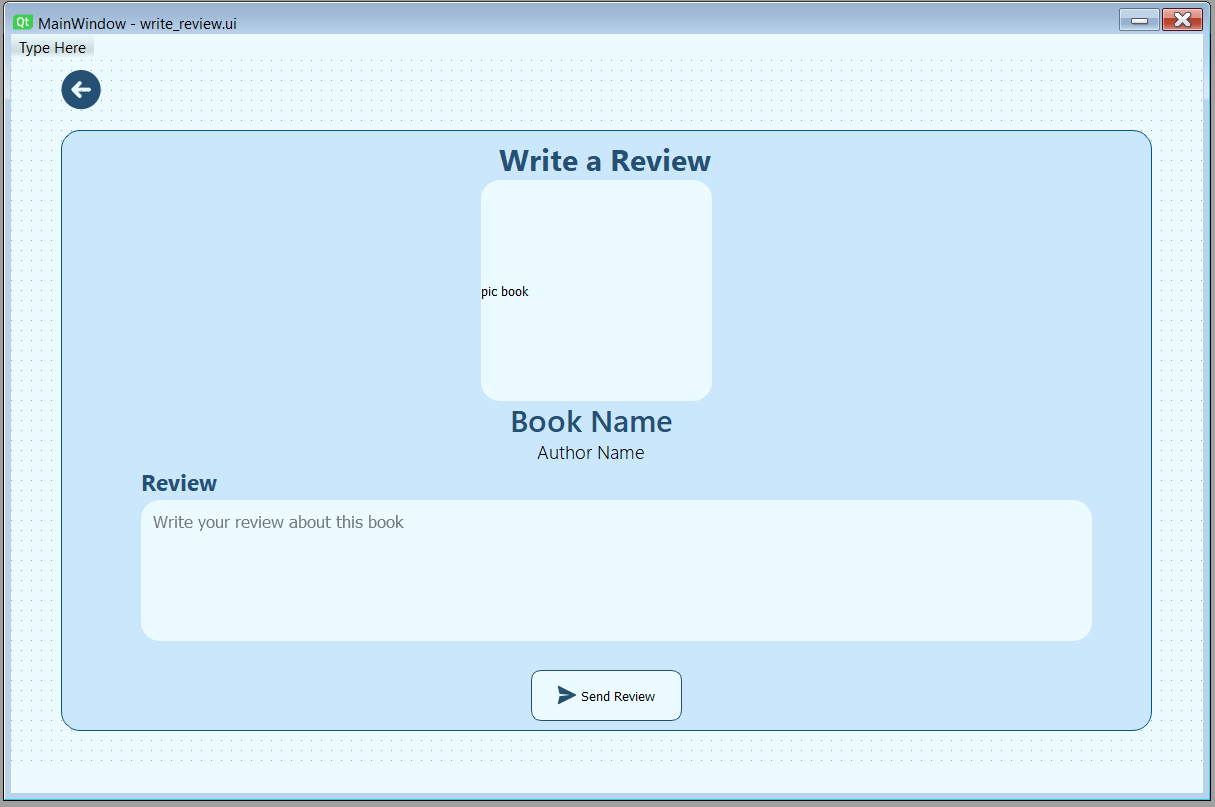
#### Figure 5.1. Bookstore homepage

Then, users can hover over the image or book title to view detailed information about the work. The book details section will display basic information such as book title, author name, genre, publisher, total number of comments, and average number of reviews. Here, users can write their own comments by clicking the “write your review” button.

****

#### Figure 5.2. Book detail page

After making a review request, the application will display a new page for customers to make a review in the textbox and send it by pressing the “Send Review” button.

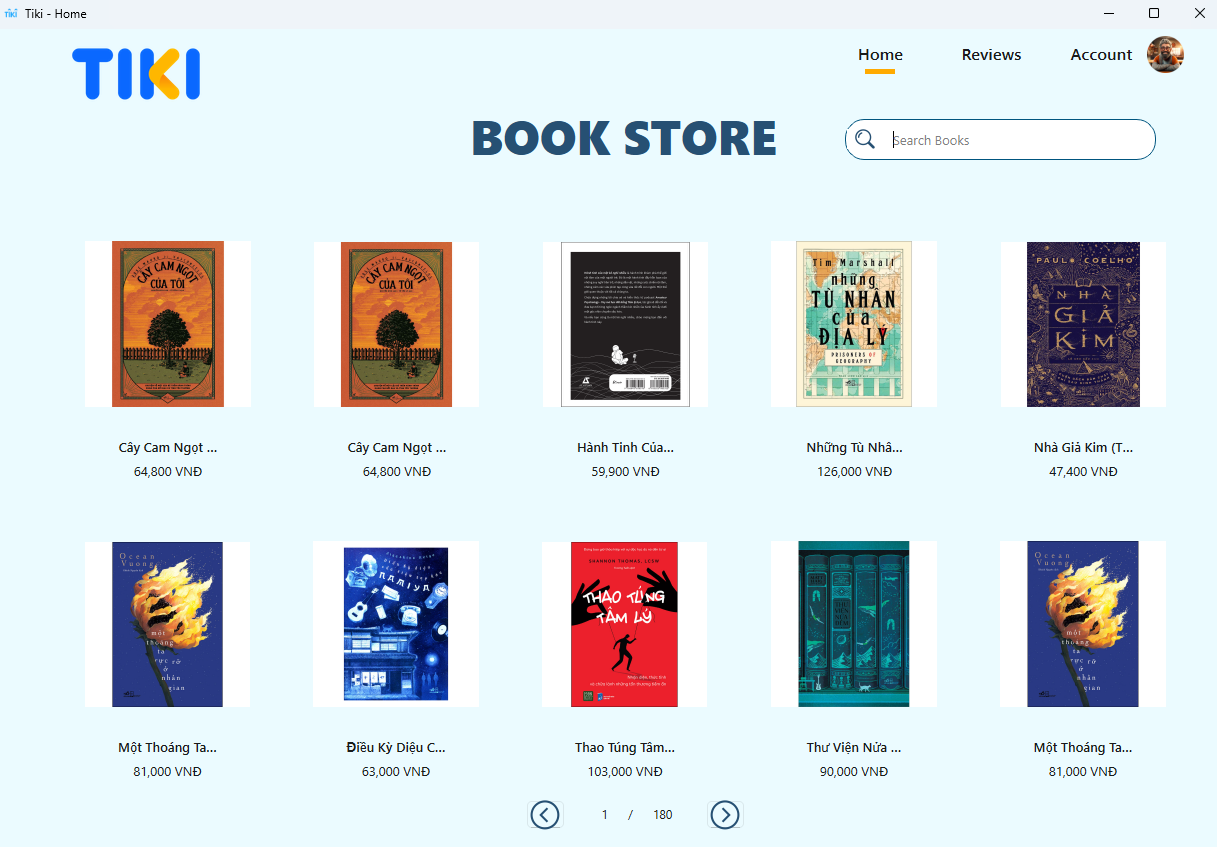
****

#### Figure 5.3. Book review writing page

## 5.2. Interface implementation

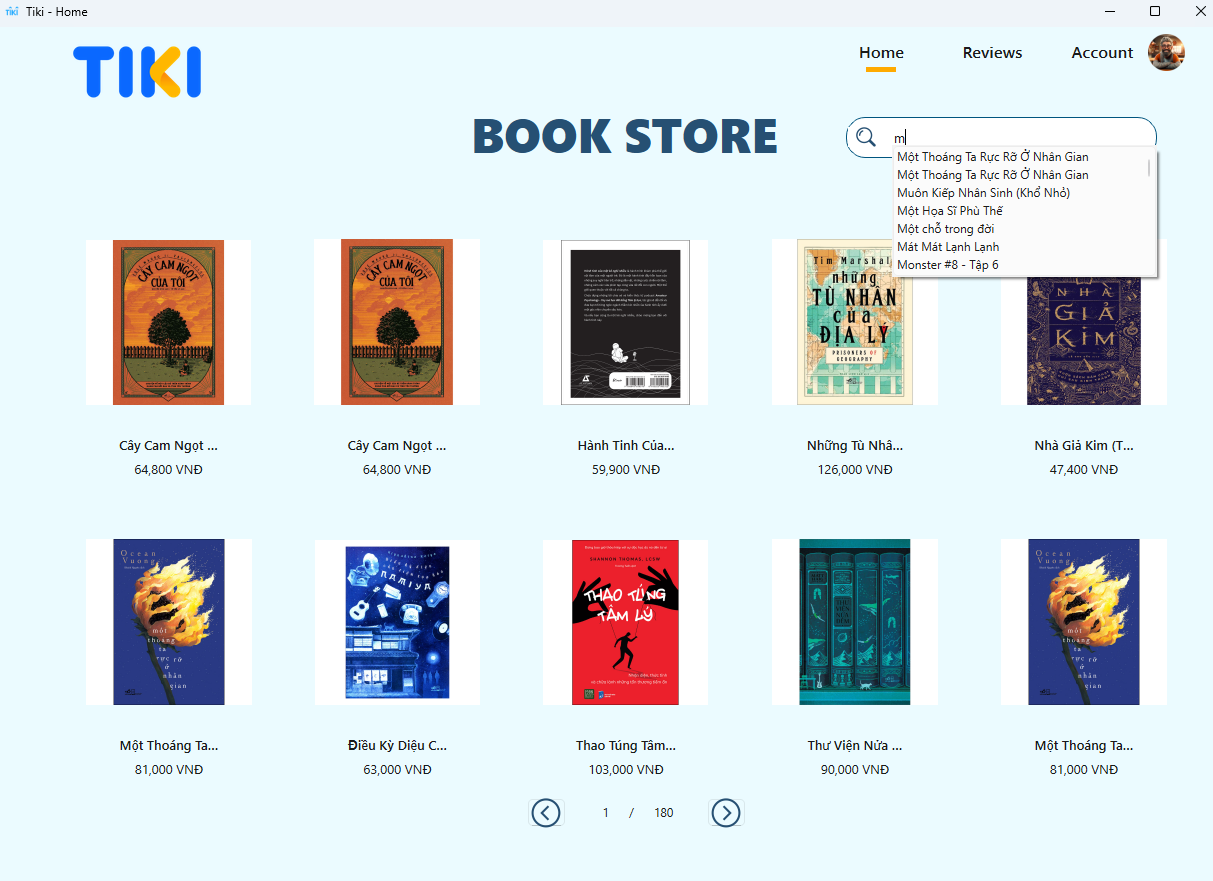
After designing the application using Qt Designer, the team implemented the software using the PyQt6 library.

### 5.2.1. Home Page



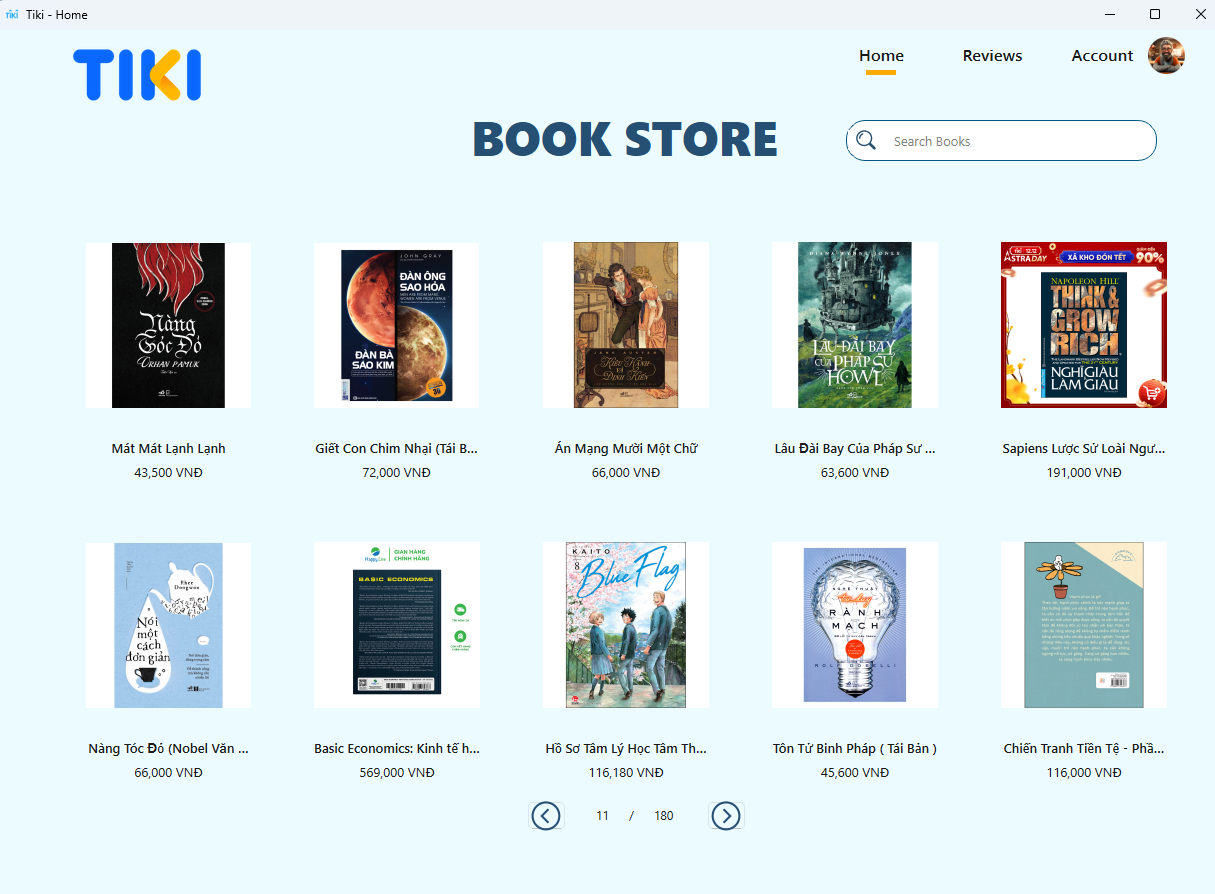
#### Figure 5.4. Home Page

In the Home interface, you can observe the books available in the data, showing the title, image, and price. Here, the user can click on the product they want to review to navigate to the product page. For the Reviews button in the top right corner, if the user clicks on it, they will be defaulted to the page of the first book in the product list, which is "Cây cam ngọt của tôi".



#### Figure 5.5. Search bar ở Home Page

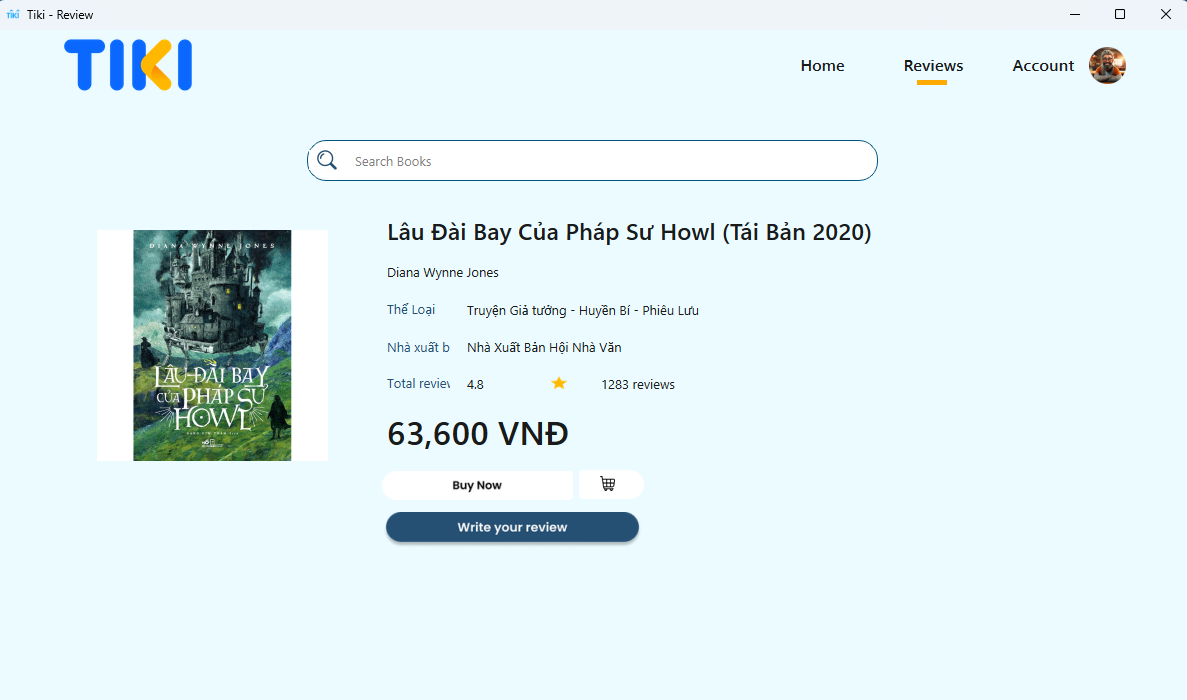
Users can search for products they want to view on the Home page, and access those products by clicking the Search button.



#### Figure 5.6. Change the product when clicking the Next or Previous button

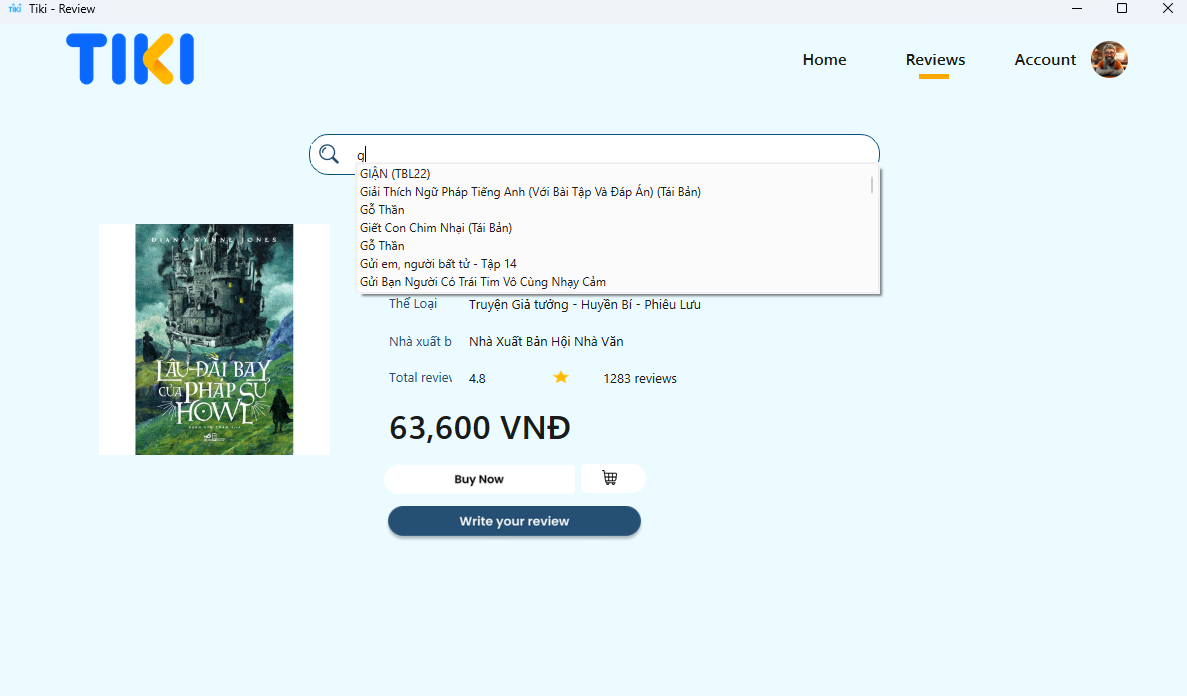
There are a lot of products, and the user can browse through 180 product pages.

### 5.2.2. Review page



#### Figure 5.7. Review Page

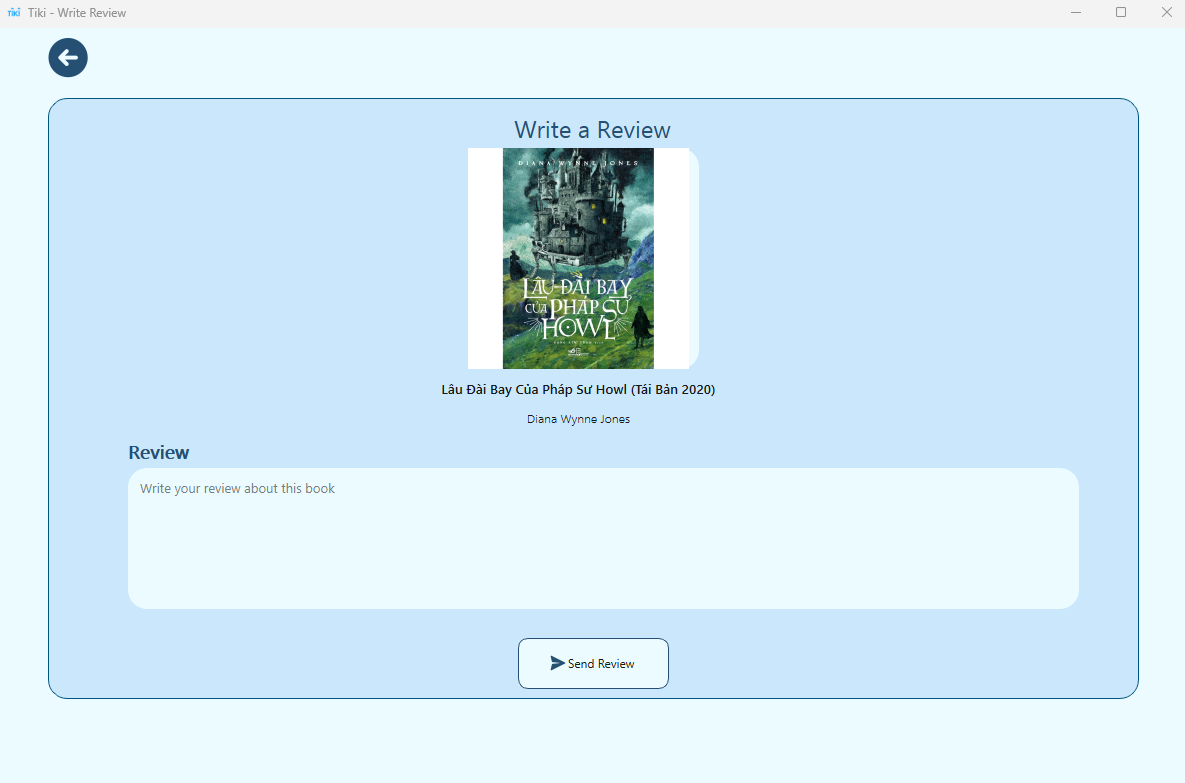
The project is designed to allow users to review products, rather than purchase them. On the Product Page, users can view information about the author, publisher, review score, and number of reviews. The Review Page also has a search bar, with similar functionality to the search bar on the Home Page.



#### Figure 5.8. Search functionality on the Review page

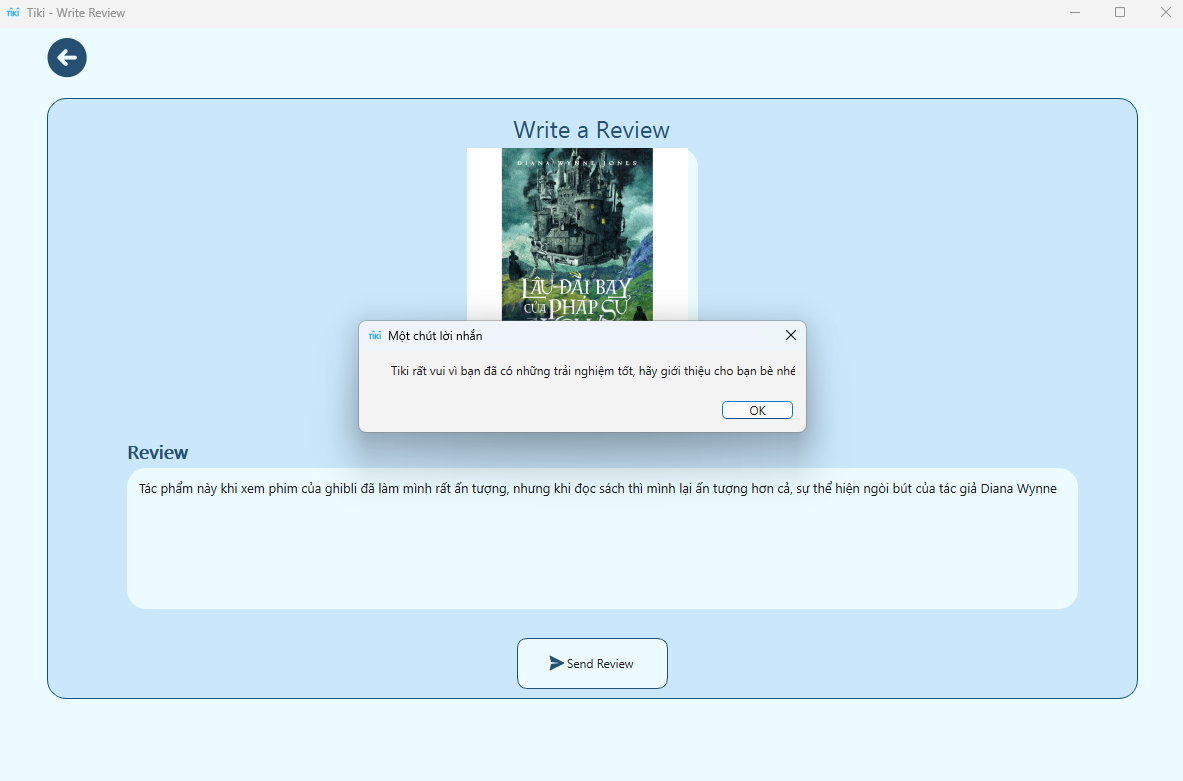
When users want to write a review for a product, they can click the "Write your review" button at the bottom to navigate to the Write Review Page.

### 5.2.3. Write Review Page



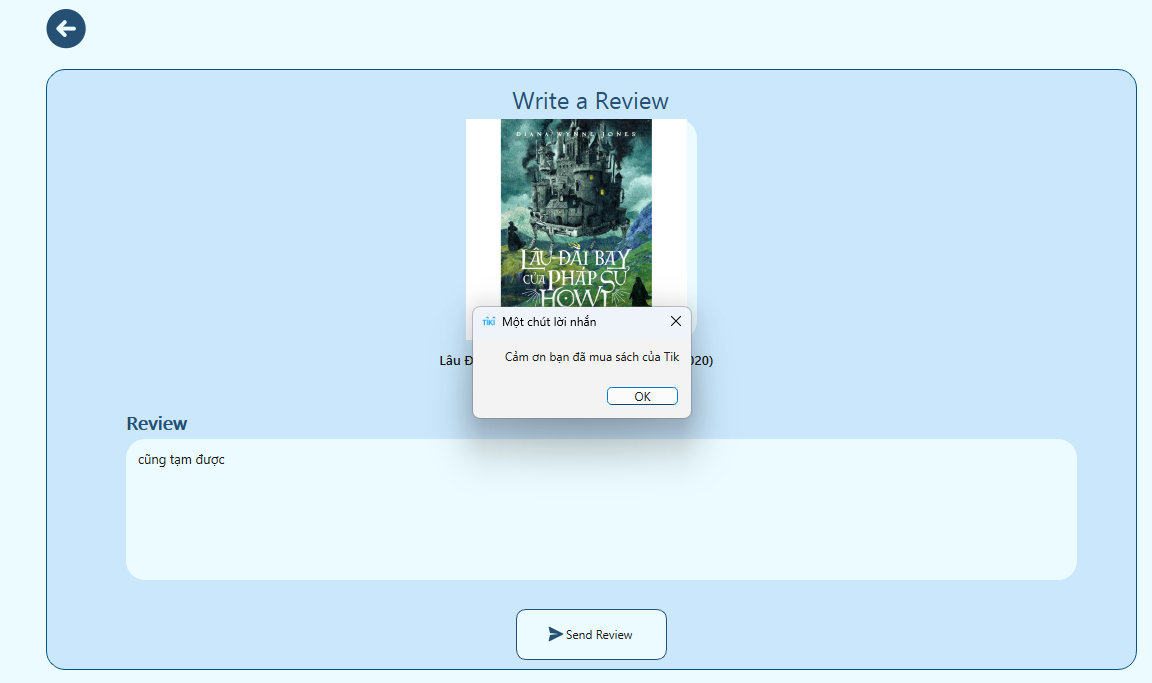
#### Figure 5.9. Write Review Page

On this page, users can write reviews for products, for example, for the book "Lâu đài bay của pháp sư Howl" by Diana Wynne Jones. A user could leave a review like this: "Tác phẩm này khi xem phim của ghibli đã làm mình rất ấn tượng, nhưng khi đọc sách thì mình lại ấn tượng hơn cả, sự thể hiện ngòi bút của tác giả Diana Wynne Jones làm mình đi từ bất ngờ này đến bất ngờ khác, mình thực sự đắm chìm vào thế giới xinh đẹp ấy.", their review will be sent and saved in the data. The application will then analyze the comment and determine if it is positive, negative, or neutral, and display the result in a message box.

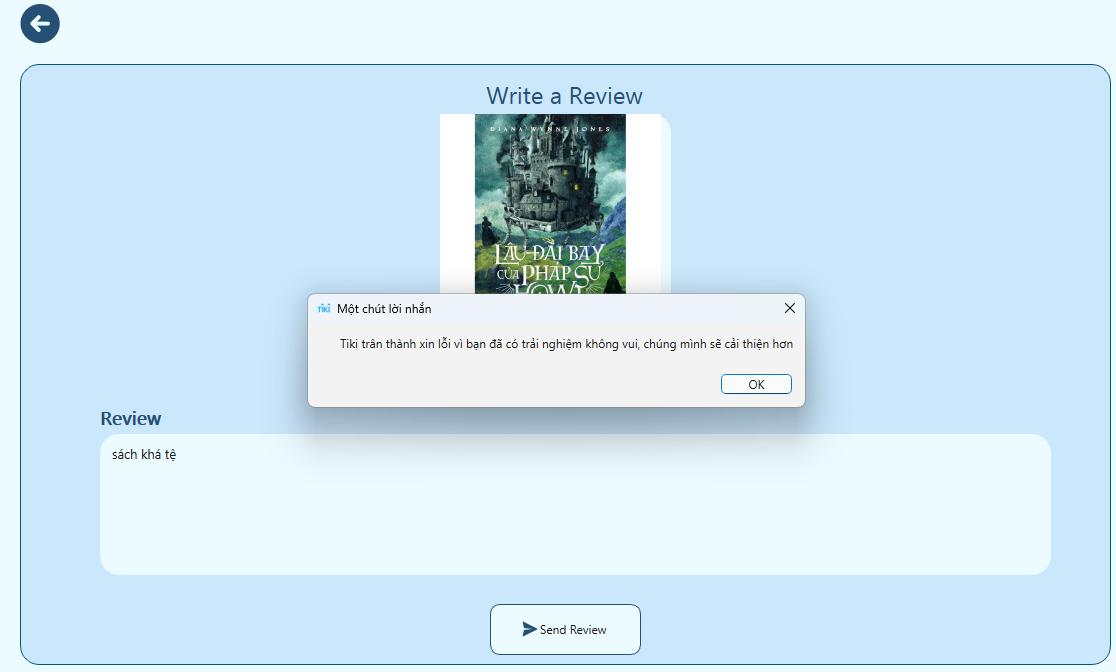


#### Figure 5.10. Message Box after the user submits their comment (Positive Message Box)

Depending on the level of positivity/negativity in the user's comment, the Message Box will display a different response.



#### Figure 5.11. MessageBox Neutral



#### Figure 5.12. MessageBox Negative

# Chapter 6: Conclusion

Chapter 6 summarizes the research findings on sentiment analysis, highlighting the effectiveness of the PhoBERT model. It also addresses limitations, such as accuracy, and suggests expanding and improving the model's application in other areas to enhance user experience.

## 6.1. Conclusion

This research has achieved important results in applying the Pho-BERT model to analyze customer emotions based on book product comments on the Tiki e-commerce platform. From the reviews posted by users on the Tiki platform, the research team successfully classified the sentiments of those reviews. The Pho-BERT model, which is optimized for Vietnamese, has allowed the research team to perform highly accurate sentiment analysis, identifying positive, negative and neutral emotions in comments. of customers. This helps create a clearer view of customer satisfaction with book products, thereby assisting managers in improving service and product quality.

In addition, the research team also successfully built an interactive user interface that allows searching for book titles and viewing related reviews, helping to increase the ability to access and use information for users. This interface not only provides an easy way to find information about books, but also displays reviews clearly and intuitively, thereby helping users make smarter shopping decisions.

These improvements not only enhance the user experience on the Tiki platform but also open up opportunities to apply sentiment analysis techniques in other fields. Results from this study provide a solid foundation for further research and practical applications, contributing to the development of sentiment analysis tools and improving user interfaces in electronics commercial environments.

## 6.2. Limit

Although the project of applying the Pho-BERT model in analyzing customer sentiment has achieved encouraging results, there are still some important limitations that need to be noted.

The size of the training dataset is quite small, which may limit the model's ability to fully recognize and analyze emotional nuances in comments. A richer and more diverse data set will help the model learn many aspects of Vietnamese emotions and vocabulary, thereby improving the accuracy and generalizability of the analysis.

The sample data used in the study may be subjective, as they are mainly based on comments from specific users on the Tiki platform. This can lead to bias in the analysis, as these comments do not necessarily fully reflect the diversity of sentiment within the broader user community. Therefore, the results of sentiment analysis may not be completely accurate and do not represent the entire emotions of customers on the e-commerce platform.

The Pho-BERT model may have difficulty handling complex semantic cases or extreme emotional situations. Such situations often require a deep understanding of context and cultural or psychological factors, which current models may not fully capture. For example, emotions in extreme situations such as strong anger or extreme happiness may require the ability to recognize specific semantic and contextual cues that the model has not been trained to recognize. exact face.

When the model is not optimized for these factors, sentiment analysis results may not reflect the true level and nuance of the user's emotions. This reduces the system's ability to provide detailed and accurate analysis of customer feedback, especially in complex or novel situations.

## 6.3. Development direction

In the future, the topic of applying the Pho-BERT model in analyzing customer emotions can be expanded in many development directions.

Improve model performance: fine-tune the algorithm or combine with other deep learning models to recognize more complex emotions

Integrate analytics across multiple product categories: from the trained model, it can be used to apply analytics to many other items on e-commerce platforms.

Combine multilingual text data analysis: in addition to Vietnamese, the research team can learn and apply emotion analysis technologies in other languages using the knowledge acquired from research. This.

These developments not only increase the value of current research but also open up opportunities to apply emotion analysis technology to many different fields and applications.

# REFERENCES

[1] Colman, B. (2024). Online review survey 2024 [statistics every marketer should know]. Customer Alliance. Retrieved August 10, 2024, from <https://www.customer-alliance.com/en/resources/article/online-review-survey-2024/>

[2] Burstein, D. (2014, August 5). Ecommerce Research Chart: Customer feedback and ecommerce success. MarketingSherpa. Retrieved August 10, 2024, from <https://www.marketingsherpa.com/article/chart/customer-experience-ecommerce>

[3] Dive into Deep Learning. (n.d.). Dive into Deep Learning — Dive into Deep Learning 1.0.3 documentation. Retrieved August 10, 2024, from <https://d2l.ai/>

[4] Deep Learning. (n.d.). Deep Learning Book. Retrieved August 10, 2024, from <https://www.deeplearningbook.org/>

[5] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2017, June 19). [1706.06083] Towards Deep Learning Models Resistant to Adversarial Attacks. arXiv. Retrieved August 10, 2024, from <https://arxiv.org/abs/1706.06083>

[6] Rane, N. (2023, April 21). (PDF) Customer experience and satisfaction: Importance of customer reviews and customer value on buying preference. ResearchGate. Retrieved August 10, 2024, from <https://www.irjmets.com/uploadedfiles/paper//issue_4_april_2023/36460/final/fin_irjmets1682057620.pdf>

[7] Eid, M. I. (2011, January). (PDF) Determinants of E-Commerce Customer Satisfaction, Trust, and Loyalty in Saudi Arabia. ResearchGate. Retrieved August 10, 2024, from <https://www.researchgate.net/publication/228867201_Determinants_of_E-Commerce_Customer_Satisfaction_Trust_and_Loyalty_in_Saudi_Arabia>

[8] Natasya, W. D. P., Musa, M. I., Haeruddin, M. I. W., Hasbiah, S., & Aslam, N. F. (2024, April 17). (PDF) Influence of Online Customer Review and Online Customer Rating On Purchasing Decision on Shopee E-Commerce: Study on Management Study Program Students, Faculty of Economics and Business, Makassar State University. ResearchGate. Retrieved August 10, 2024, from <https://www.researchgate.net/publication/379021634_Influence_of_Online_Customer_Review_and_Online_Customer_Rating_On_Purchasing_Decision_on_Shopee_E-Commerce_Study_on_Management_Study_Program_Students_Faculty_of_Economics_and_Business_Makassar_State_U>

[9] Huang, H., Zavareh, A. A., & Mustafa, M. B. (2023). Sentiment Analysis in E-Commerce Platforms: A Review of Current Techniques and Future Directions. IEEE Access, 11, 90367–90382. <https://doi.org/10.1109/access.2023.3307308>

[10] Xiong, W., Zuo, Y., Zhang, M., Zhang, C., & Guo, C. (2024). Research on Sentiment Analysis of E-commerce Live Comments based on Text Mining. Frontiers in Computing and Intelligent Systems, 6(3), 34–36. <https://doi.org/10.54097/c2wofcb2>

[11] Bằng, N. N. L., Hô, N. V., & Thành, H. T. (2021). Mô hình khai phá ý kiến và phân tích cảm xúc khách hàng trực tuyến trong ngành thực phẩm. TẠP CHÍ KHOA HỌC ĐẠI HỌC MỞ THÀNH PHỐ HỒ CHÍ MINH - KINH TẾ VÀ QUẢN TRỊ KINH DOANH, 16(1), 64–78. <https://doi.org/10.46223/hcmcoujs.econ.vi.16.1.1388.2021>

[12] Sentiment Analysis Implementing BERT-based Pre-trained Language Model for Vietnamese. (2020b, November 26). IEEE Conference Publication | IEEE Xplore. <https://ieeexplore.ieee.org/document/9335912>

[13] Tran, K. Q., Nguyen, A. T., Hoang, P. G., Luu, C. D., Do, T. H., & Kiet, V. N. (2022, June 1). Vietnamese Hate and Offensive Detection using PhoBERT-CNN and Social Media Streaming Data. arXiv.org. <https://arxiv.org/abs/2206.00524>

[14] Combining PhoBERT and SentiWordNet for Vietnamese Sentiment Analysis. (2021, November 10). IEEE Conference Publication | IEEE Xplore. <https://ieeexplore.ieee.org/document/9648599>

[15] Học sâu là gì? - Giải thích về Học sâu - AWS. (n.d.). Amazon Web Services, Inc. <https://aws.amazon.com/vi/what-is/deep-learning/>

[16] Công, N. (n.d.). Deep Learning là gì? Tìm hiểu từ A đến Z về học sâu. <https://nguyencongpc.vn/deep-learning-la-gi-tim-hieu-tu-a-den-z-ve-hoc-sau>

[17] SuNT. (n.d.). SuNT’s Blog | AI in Practical. <https://tiensu.github.io/blog/61_bert/>

[18] Khánh, P. N. (2020, May 23). Khoa học dữ liệu. Khanh’s Blog. <https://phamdinhkhanh.github.io/2020/05/23/BERTModel.html#23-masked-ml-mlm>

[19] Nguyen, D. Q., & Nguyen, A. T. (2020). PhoBERT: Pre-trained language models for Vietnamese. <https://doi.org/10.18653/v1/2020.findings-emnlp.92>

[20] Jivtode, R. (2023, April 21). NLP: Natural Language Processing - Rani Jivtode - Medium. Medium. <https://medium.com/@rani.jivtode21/nlp-natural-language-processing-8667615f4d58>

[21] Kapadia, B., & Jain, A. (2021). Analysis of Papers Based on Sentiment Analysis Applications on E-Commerce Data. In Advances in intelligent systems and computing (pp. 323–340). <https://doi.org/10.1007/978-3-030-73603-3_30>

[22] Tiki Books Dataset. (2023, January 1). Kaggle. <https://www.kaggle.com/datasets/biminhc/tiki-books-dataset/data>

[23] Duyệt, L. (n.d.). GitHub - stopwords/vietnamese-stopwords: Vietnamese stopwords. GitHub. Retrieved August 10, 2024, from <https://github.com/stopwords/vietnamese-stopwords>