VIETNAM NATIONAL UNIVERSITY
HANOI INTERNATIONAL SCHOOL

Subject: ADVANCED DATA ANALYTICS

Report Final Examination

Class	:	INS3075
Lecture	:	TS Truong Cong Doan
Topic	:	Analysis customer sentiment by Foody.vn
Group number	:	2
Members	:	Do Thi Minh Phuong – 21070074
		Nguyen Phuong Thao – 20070981
		Nguyen Anh Đuc – 20070921
		Be Hang Nga – 20070961

Hanoi, 17th January 2025

Table of Contents

List of figure

List of table

I. INTRODUCTION

1.1 Problem Statement

The rapid growth of e-commerce and social networking platforms has transformed how consumers share their experiences and opinions on products and services. In the food industry, online reviews have become a crucial source of customer feedback, allowing businesses to gain insights into customer satisfaction, preferences, and expectations. In particular, the online food ordering industry has seen a surge in platforms such as Foody.vn, where customers frequently leave reviews about their experiences with various restaurants and food delivery services. These reviews contain valuable sentiment data, reflecting customer perceptions regarding food quality, delivery speed, pricing, and service experience. However, due to the large volume of reviews generated daily, manually analyzing customer sentiment is impractical. This has created a demand for automated sentiment analysis tools, which can efficiently process and classify thousands of customer reviews. Machine learning and natural language processing (NLP) techniques offer a powerful solution to this problem, enabling businesses to extract meaningful insights from customer feedback

1.2 Objective

This study aims to classify customer sentiments as either positive or negative by analyzing a dataset of 33,633 food ordering service reviews collected from Foody.vn during the period 2016 –2021. By applying machine learning models, the objective is to determine the most accurate approach for sentiment classification. The model with the best performance will be used to analyze and predict customer sentiment in new reviews, allowing businesses to understand how customers perceive their services in real-time. Through this study, business managers will gain data-driven insights into customer emotions, helping them make informed decisions regarding service improvements, marketing strategies, and customer relationship management. Ultimately, the goal is to optimize business operations, enhance customer experience, and increase competitiveness in the online food delivery market.

1.3 Related Paper

1.3.1 Sentiment Analysis of Customer Feedback in Online Food Ordering Services

Advantages:

- **Comprehensive Dataset:** This study analyzed a substantial dataset of **236,867 Vietnamese online reviews** from platforms like **Foody.vn** and **Diadiemanuong.com**, spanning **2011 to 2020**, providing a broad temporal perspective on customer sentiments.
- **High Accuracy:** The proposed method achieved an impressive **accuracy of up to 91.5%**, indicating the effectiveness of their sentiment analysis approach.

Limitations:

- **Language Specificity:** The focus on Vietnamese reviews may limit the generalizability of the findings to non-Vietnamese contexts.
- **Platform Limitation:** Data was sourced from only two platforms, which might not capture sentiments from users of other popular food ordering services.

1.3.2 Sentiment Analysis of Online Food Reviews using Customer Ratings

Advantages:

- **Rating-Based Classification:** This research proposed a system that classifies reviews on a **scale of 1 to 5** based on sentiment, offering a nuanced understanding beyond binary positive/negative classifications.
- **Visualization:** The study utilized **word clouds** to display groups of words influencing ratings, aiding in the intuitive understanding of sentiment drivers.

Limitations:

- **Limited Dataset:** The experimental analysis was conducted on the **latest 3,000 fine food reviews**, which may not be representative of the broader spectrum of customer opinions.
- **Scalability Concerns:** The approach's effectiveness on larger datasets remains untested, raising questions about its scalability.

II. DATA SOURCE

2.1 Data Crawling

Beautiful Soup and Selenium libraries in Python language are used to collect data on the website. Data collection is based on the Hypertext Markup Language (HTML) structure of the Foody.vn site. If you want to collect data of any information, retrieve the data corresponding to the HTML tag containing that information.

- Coding crawling data and connecting MongoDB

Step 1: Define a Web Scraping Class (import libraries, define crawlShop Class, Crawl data method)

Step 2: Set up Selenium for Dynamic Content (initialize WebDrive, Scrolling Function, Extract elements)

Step 3: Loop Through Multiple Locations (Define Location Dictionary, Iterate Over Locations, Extract Business Details)

Step 4: Set up MongoDB Connection (Define MongoDB Connection Class, Connect to Database)

Step 5: Defining the Comments Scraping Class (crawlComments Class, WebDriver Setup, Scrape Review, and Extract Data)

Step 6: MongoDB Integration and display data in MongoDB

- Configure Docker Compose and create Docker Container for MongoDB

Step 1: Define a Web Scraping Class

- Import Required Libraries: Includes requests, BeautifulSoup for HTML parsing, and type hinting.
- Define crawlShop Class: This class takes a list of URLs to crawl and fetches data.
- Crawl Data Method: The crawl_data() method sends HTTP requests to each URL, parses the HTML response, and prints the parsed content (potentially for further processing).

Step 2: Set Up Selenium for Dynamic Content

- Import Selenium Libraries: Necessary modules for web automation are imported.
- Initialize WebDriver: Uses Selenium's Chrome WebDriver to access a specified webpage.
- Scrolling Function: Defines scroll_down_page() to scroll down the webpage and load dynamic content.
- Extract Elements: Uses WebDriver to wait for and extract specific elements by class name, printing the text content for each.

Step 3: Loop Through Multiple Locations

- Define Location Dictionary: Maps URL-friendly location names to display names.
- Iterate Over Locations: For each location, accesses the location page, scrolls, and extracts data.
- Extract Business Details: Retrieves name, link, address, and location for each business, printing them.

Step 4: Set Up MongoDB Connection

- Define MongoDB Connection Class: Uses pymongo.MongoClient to connect to a MongoDB database.

- Connect to Database: Establishes a connection to MongoDB and confirms successful connection.

Step 5: Defining the Comments Scraping Class

- crawlComments Class: Manages comments scraping.
- WebDriver Setup: createDriver() opens Chrome; deleteDriver() closes it.
- Scrape Reviews (scrape_reviews_and_comments_from_link): Open URL, click "Load More" repeatedly, scroll to load all content and parse HTML with BeautifulSoup.
- Extract Data: Collects Id, shop_id, name, review_date, rating, review_text, user_name for each review.
- Return Data: Outputs a list of review dictionaries.

Step 6: MongoDB Integration and display data in MongoDB

- Integrate with MongoDB: After scraping each review, insert it into the comments collection with fields such as Id, shop_id, name, review_date, rating, review_text, user_name.
- View Data: Use MongoDB commands (e.g., find()) to verify that data is correctly inserted and view documents.

2.2 Data description

The collected dataset had 33,632 records, including ID user, ID shop, review date, review text, user name, and rating of a for that store. The number of reviews gathered from foody.vn is 33,632 comments. This dataset will go into the preprocessing and cleaning step to provide input to the later steps of the models.

We combined the dataset and split it into training and test sets with shuffling at the ratio of 70%:30%. The original combined dataset contains 33632 pieces of data, and Table 1. shows the distribution of data in the training and test sets.

Training	Test
22073	11559

Table 1. Distribution of data

III. EXPLORATORY DATA ANALYSIS (EDA)

3.1 Data Cleaning

Remove blank/NULL and duplicate data: the collected dataset will have no null value and 2098 duplicated data which does not make sense in the analysis process, causing a waste of storage memory so we will drop them.

3.2 Data Processing

Text Processing: Phân tích và xử lý dữ liệu dạng văn bản (tokenization, stemming, lemmatization, stop-word removal). Use a Vietnamese language model (vi_core_news_lg) in spaCy to create a scatter text visualization. The apply function applies the clean_text() method to each entry in the review_text column, and passes several parameters to this method: emoji_dict, teen_dict, wrong_lst, and stopwords_lst:

- Emoji Conversion: Replaces or processes emojis based on the emoji_dict.
- Teen Code Translation: Translates teen slang to their expanded form using teen_dict.
- Wrong Word Correction: Replaces any incorrect words from the wrong_lst.
- Stopword Removal: Removes stop words listed in stopwords_lst to focus on the key content of the text.

Comments after cleaning:

review_text	review_text_clean
Gà tắm mắm, phô mai kéo sợi siêu ngon, giá mềm	gà tắm mắm phô mai kéo sợi siêu ngon giá mềm
Gà BBQ Hàn, phô mai kéo sợi siêu ngon, giá mềm	gà bq hàn phô mai kéo sợi siêu ngon giá mềm
Gà với khoai tây quá mặn	gà với khoai tây quá mặn
Mình vừa đặt 1 phần gà 92k và vô cùng thất vọn...	mình vừa đặt 1 phần gà 92k và vô cùng thất vọn...
Đồ ăn chuẩn vị hàn quốc, ngon giá cả hợp lí	đồ ăn chuẩn vị hàn quốc ngon giá cả hợp lí
...	...
Bánh canh ngon, ăn ko ngán, bánh lọc thì tuyệt...	bánh canh ngon ăn không ngán bánh lọc thì tu...
Được người bạn giới thiệu vào đây, các bạn có ...	được người bạn giới thiệu vào đây các bạn có ...
Quán nằm trong đường Trần Quang Diệu, mà phải ...	quán nằm trong đường trần quang diệu mà phải ...
Trời mưa lạnh thèm ăn bún bò, lần đầu lại đây ...	trời mưa lạnh thèm ăn bún bò lần đầu lại đây ...
Quán chủ yếu là phở nhưng mình lại thích bún b...	quán chủ yếu là phở nhưng mình lại thích bún b...

Figure 1. Comments after cleaning

3.3 Data labelling

It can be seen that most of the scores fall within the range of (7,8), with most reviews rating quite well, while only a few reviews rate the restaurant as average (5) and low:

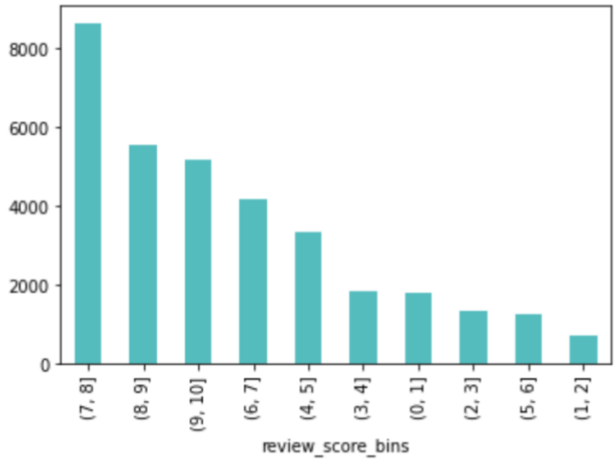

Figure 2. Distribution of Review Scores

To perform the data labeling process before being trained, the research applied the classifying emotions method according to the customer rating to divide the collected dataset into 2 datasets, labeled according to the following rules:

- Rate < 5: Reviews below 5 stars will be labeled negative.
- Rate >= 5: Review comments rated above 5 stars will be labeled as positive.

Type	Number of reviews	% of total
Negative	8013	17.3%
Positive	23521	82.7%
Total	31534	100%

Table 2. Distribution of Labeled Positive and Negative Reviews

3.4 Feature extraction

In machine learning, computers cannot understand natural languages directly but can only understand languages when they are represented in the form of vector space. The input attribute dimensions will be represented as a vector matrix. There are many methods to represent text into a vector matrix, for example: traditional ways such as the Bag of N-grams model, TF-IDF model, topic models or improved ways such as Word2Vec, GloVe, FastTex models (Sarkar, 2019). In this study, we apply two methods, Bag of N-grams and TF-IDF, to test the model and represent the data:

Bag of word (BoW)

The BoW model only gathers all words in the form of a single word, and does not contain phrases consisting of many words combined. The Bag of N-Grams model will solve this problem. Bag of N-grams will form a set of phrases consisting of n-words that can be grouped together depending on needs. For example, consider the following three comments: “the food is so good”, “the service is so bad”, “the food is not good”.

TF-IDF

Bag of word n-grams model has some problems for large data sets, that is, words have a high frequency in most text paragraphs, but have no distinguishing meaning. categories, such as the words "this", "that", "very", "store", etc. Then the TF-IDF index will be used to calculate and detect words with high and low weights.

- TF (Term Frequency): The number of times a word appears in a document is its Term Frequency. A higher value means a term appears more often than others, and so, the document is a good match when the term is part of the search terms.
- IDF (Inverse Document Frequency): Words that occur many times in a document, but also occur many times in many others, may be irrelevant. IDF is a measure of how significant a term is in the entire corpus.

IV. METHODOLOGY

4.1 Overview model and methods

The research data was collected for research purposes, containing raw data from the Foody.vn websites. Before the machine learning procedure, the raw data is preprocessed, sampled, and labeled. The training dataset is used during the learning process and is used to fit the parameters and tune the hyperparameters of a classifier. Test datasets are used only once as the final step to reporting estimated error rates for future predictions. Figure 2 is an overview of the research model which we have done:

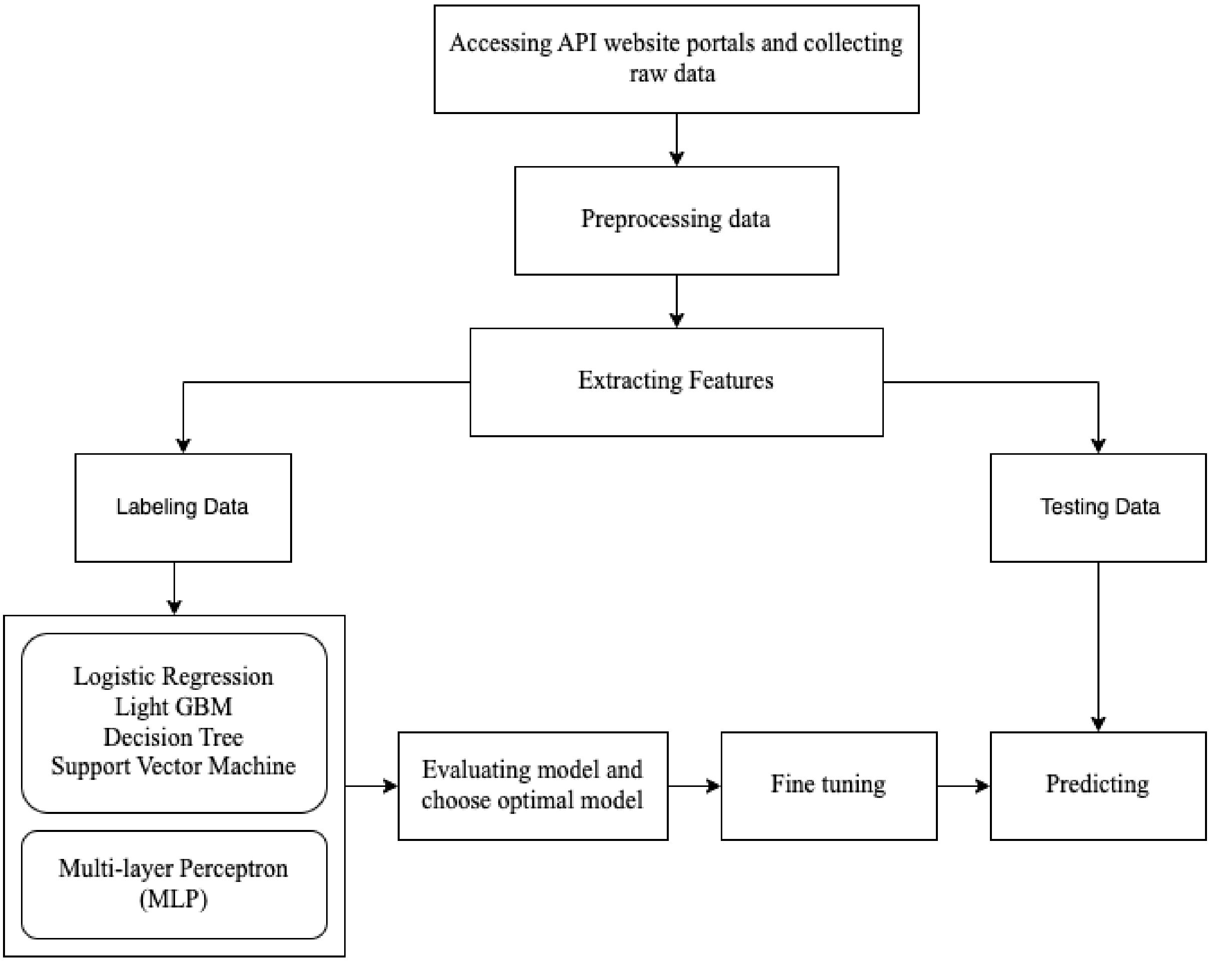

Figure 3. Proposed Overview Model and Methods

4.2 Model Implementation

This is the most important stage to determine whether a customer comment is "positive" or "negative". Based on the results of the previous research related to the topic, find the most suitable model for the dataset, which is the classified comments. Then, forecasting the unsorted comment data or new comment data arises without retraining.

4.2.1 Unbalanced data

The choice of Logistic Regression, LightGBM, Random Forest, SVM, and MLP ensures a well-rounded approach to imbalanced review sentiment analysis. Each model contributes unique strengths, making the combination effective for achieving high accuracy, recall, and precision while addressing class imbalance challenges. This strategy helps ensure that both majority and minority classes are represented fairly in the final predictions.

Result on unbalanced data:

Model	Accuracy	Precision	Recall	F1-Score
LR	0.8850	0.9514	0.8921	0.9208
Light LGBM	0.8828	0.9406	0.9004	0.9201
RF	0.8700	0.8633	0.9819	0.9188
SVM	0.8850	0.9455	0.8983	0.9213
MLP	0.8708	0.8800	0.8700	0.8700

Table 3. Result on unbalanced data

The models have high accuracy on both the training and test sets (>80%), indicating that they are performing well. However, if we look at the F1-score for class 0, we see that the results are not optimal. This is a common sign when working with imbalanced data: the model tends to predict better for the majority class (class 1 in this case), and this affects the F1-score of the minority class (class 0).

Due to the imbalanced data leading to poor prediction for class 0, SMOTE needs to be applied to balance the ratio between classes. We then build a new model on the balanced dataset to hopefully improve the accuracy, recall, and F1-score for both classes, especially for the minority class (class 0). SMOTE has generated additional samples for class 0 so that the number of samples for class 0 and class 1 is equal (both 16432 samples).

4.2.2 Balanced data

Result on balanced data after SMOTE:

Model	Accuracy	Precision	Recall	F1-Score
LR	0.8856	0.9384	0.9069	0.9069
Light LGBM	0.8860	0.9240	0.9238	0.9239
RF	0.8817	0.8911	0.9594	0.9240
SVM	0.8906	0.9363	0.9163	0.9262
MLP	0.8727	0.8700	0.8700	0.8700

Table 4. Result on balanced data after SMOTE

Linear Regression:

- Test Accuracy: The test accuracy is almost the same (before SMOTE: 88.50%, after SMOTE: 88.56%).
- Recall: After SMOTE, the recall for the positive class (1) improved from 0.8921 to 0.9069, meaning the model is better at detecting the minority class.
- Precision: The precision for the positive class (1) slightly decreased after SMOTE (from 0.9514 to 0.9384), indicating a slight increase in false positives.
- F1-Score: The F1-score remained almost unchanged (0.9208 before, 0.9224 after), reflecting a balanced performance.
- Confusion Matrix: After SMOTE, the model has fewer false positives (422 vs. 323) and fewer false negatives (660 vs. 765), indicating that resampling has improved performance on the minority class.

LGBM:

- Test Accuracy: There is a slight improvement in the test accuracy after SMOTE (from 88.28% to 88.60%).
- Recall: After SMOTE, recall for the positive class (1) increased from 0.9004 to 0.9238, showing better detection of the minority class.
- Precision: Precision for the positive class (1) slightly increased from 0.9406 to 0.9240.
- F1-Score: The F1-score improved from 0.9201 to 0.9239, reflecting a better balance between precision and recall.
- Confusion Matrix: After SMOTE, the model has fewer false positives (539 vs. 403) and false negatives (540 vs. 706), indicating better handling of the minority class after resampling.

The data balance improves the accuracy and recall of class 0 across all five models, helping the model to better recognize the minority class, showing a balance between recall and precision

V. MODEL EVALUATION

The data set has been trained with machine learning and neural network models, using algorithms including: Logistic Regression, Light GBM, Decision Tree, SVM and MLP. To evaluate the performance of the models constructed, we used the most commonly used metrics:

Class designation		**Actual Class**	
		True (1)	False (0)
Predicted Class	Positive (1)	TP	FP
	Negative (0)	FN	TN

Table 5. Confusion matrix

- **Accuracy** is calculated as the sum of two accurate predictions (TP + TN) divided by the total number of data sets (P + N). The best accuracy is 1.0, and the worst is 0.00

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} = \frac{TP + TN}{P + N}$$

- **Precision** is calculated as the number of correct positive predictions (TP), divided by the total number of positive predictions (TP + FP). The best accuracy is 1.0 and the worst 0.0.

$$Precision = \frac{TP}{TP + FP}$$

- **Recall** is calculated as the number of correct positive predictions (TP), divided by the total number of actual positive instances (TP + FN). The best recall is 1.0 and the worst is 0.0.

$$Recall = \frac{TP}{TP + FN}$$

- **F-Measure** or **F-score** is a measure of the accuracy of the test. It is calculated, based on precision and reminders, by the formula:

$$F\text{-}Score = \frac{2 \cdot precision \cdot recall}{precision + recall}$$

These 4 metrics are the most widely used in the world of machine learning, especially for classification problems. It allows us to evaluate the performance of a classifier from different perspectives. Normally, 'Accuracy' is the most representative metric for the evaluation because it reflects the situation of classification completely.

Overall, SVM is the best model.

VI. MODEL IMPROVEMENT

6.1 Ensemble Learning (Stacking)

Using Ensemble learning (stacking) helps improve the model's generalization ability, allowing the model to learn more important patterns from balanced datasets, while also reducing bias and variance. (variance). Ensemble helps increase the recall of this class due to the consensus of the models when making predictions, thereby optimizing overall accuracy and F1-score.

Set these three models (Logistic Regression, Light GBM, Decision tree and SVM) as base learners in a StackingClassifier. Then use a stronger model, such as Logistic Regression, as the final estimator.

```
Accuracy: 0.8944
Confusion Matrix:
[[1698  674]
 [ 325 6764]]
Classification Report:
              precision    recall  f1-score   support

           0       0.84      0.72      0.77      2372
           1       0.91      0.95      0.93      7089

    accuracy                           0.89      9461
   macro avg       0.87      0.84      0.85      9461
weighted avg       0.89      0.89      0.89      9461
```

Figure 4. Classification Report of StackingClassifier

The stacking model achieved a fairly good accuracy (89.44%) and also had very good precision, recall, and F1-score, especially in identifying the minority class.
We still can improve the performance of the stacking model through hyperparameter tuning using random search: Searching for optimal parameters for the model.

6.2 Testing

Optimize hyperparameters for both base and meta-models. The performance of the stacked ensemble can be greatly influenced by hyperparameter adjustment. Experiment with various hyperparameter settings for both the base models and the meta-models to find the best combination for maximizing prediction performance.

Hyperparameter Tuning Steps:

Step 1: Define Your Base Models and Meta-model

In a stacking model, you have base models (like LightGBM, RandomForest, SVM) and a meta-model (like LogisticRegression). You need to adjust the hyperparameters of both base models and meta-models.

Step 2: Set Up Hyperparameter Grid for Each Model

Each model in the stack has its own set of hyperparameters that can be tuned.

Step 3: Apply Grid Search

Result:

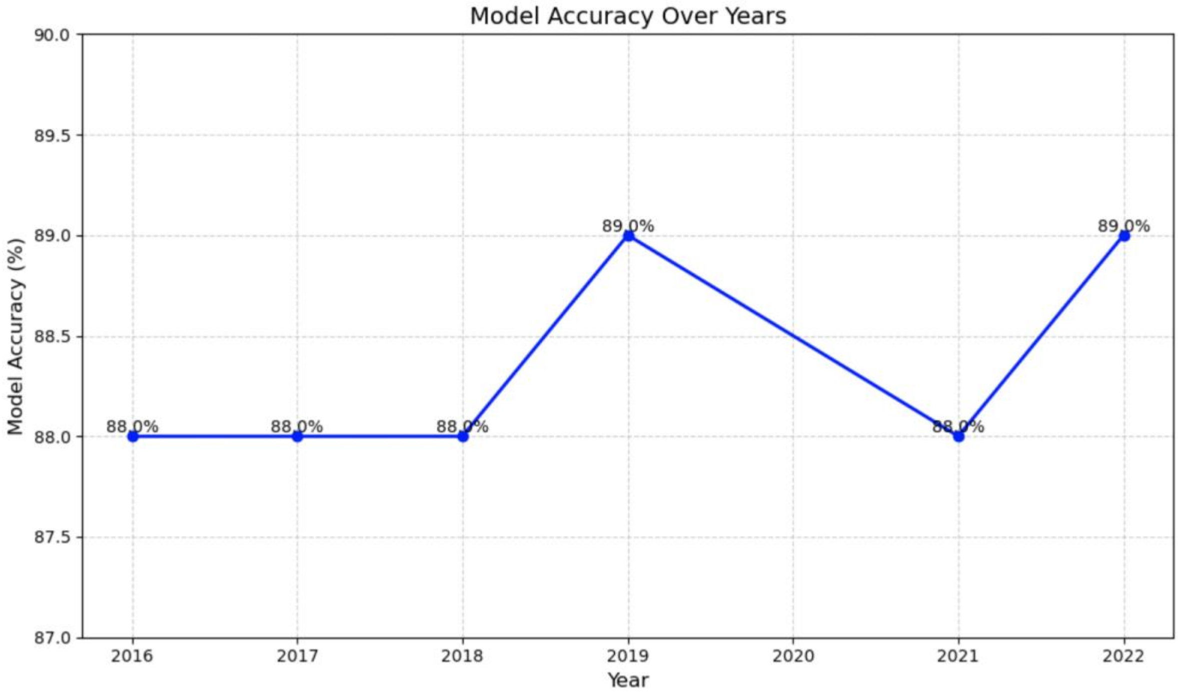

Figure 5. Model Accuracy by year (2016-2022)

VII. VISUALIZATION

Problem Statement:

1. Which restaurants have the highest average ratings?
2. Which restaurant has the highest percentage of positive reviews?
3. Which restaurant has the highest percentage of negative reviews?
4. How has the overall sentiment (positive or negative) of reviews changed over the years?
5. Which cities have the highest percentage of positive and negative reviews?

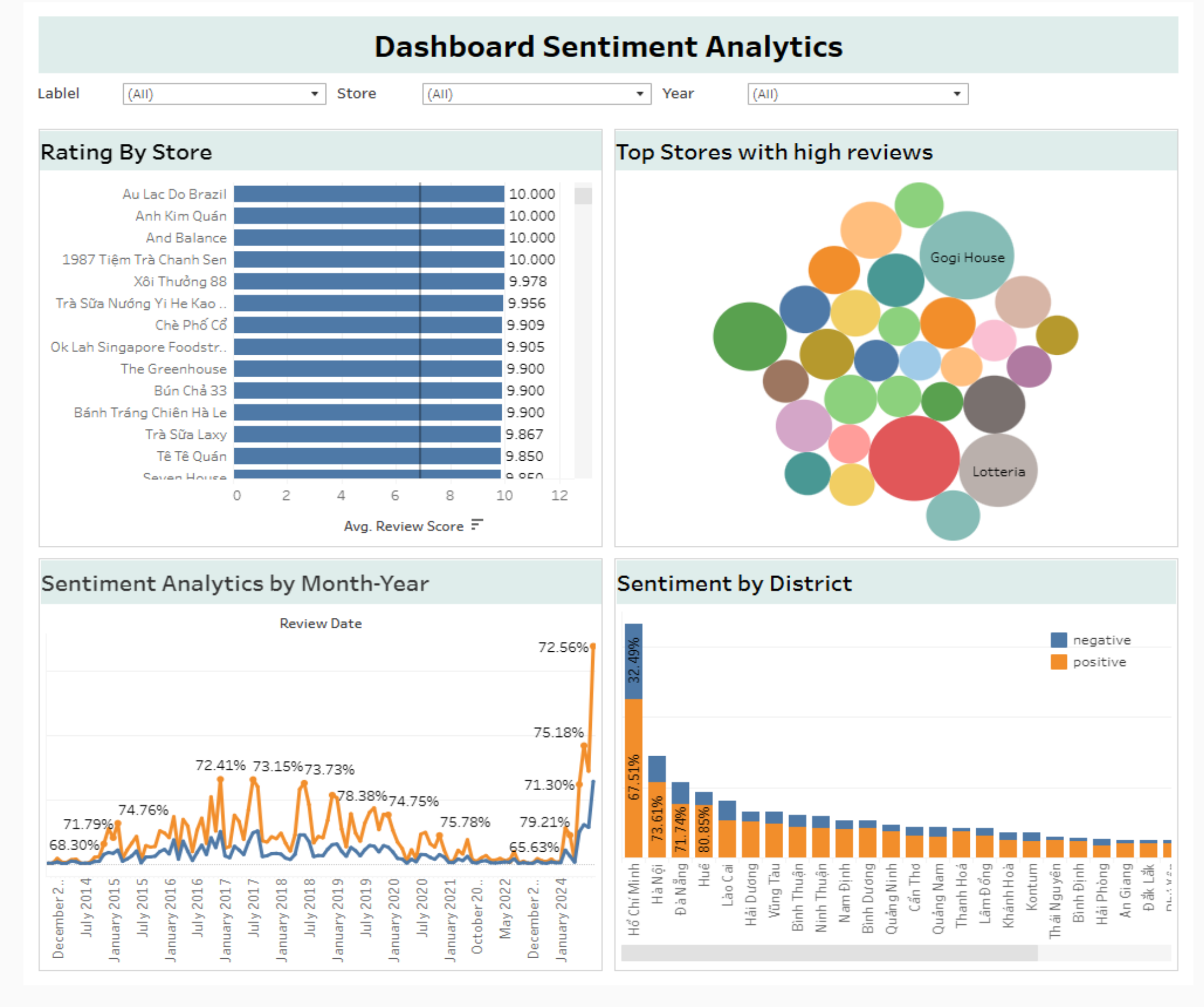

Figure 6. Dashboard Sentiment Analytics

The visualization results in the Figure above include the following four charts:
Ranking by Store, Top Rated Stores, Sentiment Analytics by Month-Year, and Sentiment by City Area.
In the first **Rating by Store** chart, we have listed the restaurants with the highest average rating.
In the **Top Rated Stores** chart, we have filtered out the stores with the best Review Score (Some highly rated restaurant chains such as Gogi House, Lotteria, Highland…).

In **Sentiment Analytics by Month-Year**, we compare the positive and negative reviews received over the months of the year in the data set, for example, in January 2024, 72.56% were good reviews - 27.46% were negative reviews.
In the last chart, we have compared the positive and negative reviews received by region. According to the assessment, the two major cities of Ho Chi Minh and Hanoi have the highest rate of assessment (Ho Chi Minh 67.51% - positive, 32.49% - negative / Hanoi 73.61% - positive, 26.39% - negative)

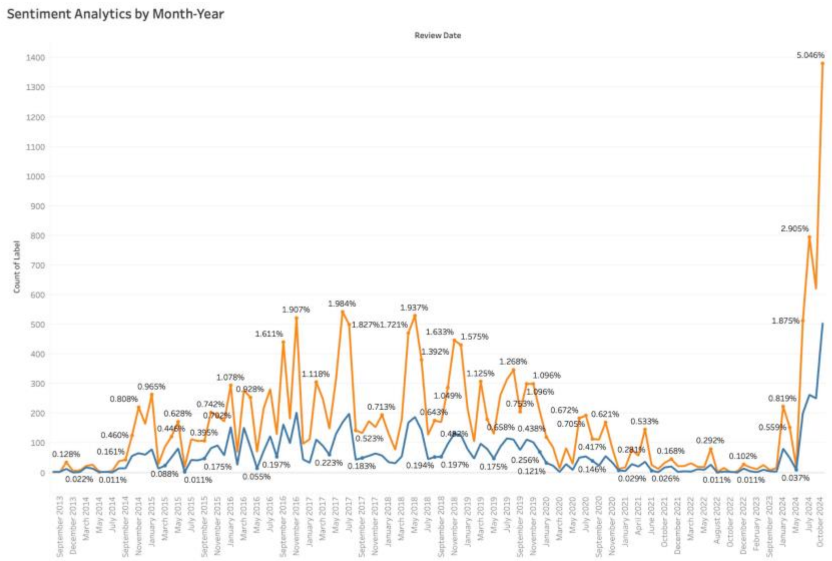

Figure 7. Sentiment Analytics by Month-year

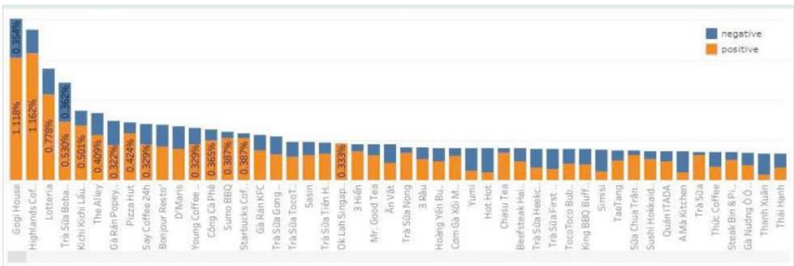

Figure 8. Sentiment by District

We look at the positive and negative reviews across the entire dataset we collected. With these two charts, we prioritize by month of the year and restaurant systems. For example: In early October 2024, according to the collected dataset, 5.046% of the positive reviews were calculated across the entire input data, while negative reviews were approximately 1% Next, according to the restaurant review system that we aggregated across multiple regions - with a restaurant system with a high rating in the dataset like Gogi House, they accounted for 1.11% of positive reviews and 0.354% of negative reviews calculated across the entire dataset we collected

Figure 9. Vietnamese Word Cloud by Positive and Negative

From the word cloud, we can see the difference between the frequent words of the two classes:
1 (positive): having key words with positive meanings such as *"ngon, mềm, chuẩn, hợp lý…"*
0 (negative): having key words with negative meanings such as *"mặn, không_có, không tiện, thất vọng,…"*

VIII. CONCLUSION

These results of the study bring into light the importance of sentiment analysis to understand customer feedback and improve business decisions in online food delivery services. Business companies can make use of machine learning models for automatically classifying customers' sentiments in an effective way that will enable them to:

- Improve the quality of the service by identifying key drivers of negative experiences.
- Optimize marketing strategies by focusing on elements that contribute to customer satisfaction.
- Refine business operations to include price adjustments, enhancement of services, and improvement in delivery.

IX. LIMITATIONS & RECOMMENDATION

9.1 Limitations

Despite its promising results, this study has several limitations. First, it was gathered from a single place, Foody.vn, which in itself may or may not be representative of customer sentiments. While this study has applied machine learning models to merely two classes of sentiment classification, positive and negative, it cannot identify the more detailed aspects of customer feedback. For example, it will not say whether the review is negative in relation to the quality of food, delivery service, price, or ambiance inside the restaurant. In fact, because it does not classify such things, it will limit the depth of the business intelligence that would have been obtained.

Also, another challenge may be the complication in text pre-processing in Vietnamese like slang words, abbreviations, misspellings. Although some methods of preprocessing have been done, linguistic variations can still affect model performance. Lastly, the model may not easily classify data they haven't seen, especially those where the patterns of customer language vary with time. This is if the model may become outdated because it is not regularly updated.

9.2 Recommendations

To enhance future research, several improvements can be made. First, expanding the dataset to include multiple food review platforms would provide a more comprehensive view of customer sentiment. Additionally, incorporating real-time data updates would allow businesses to track sentiment trends more accurately and adapt to evolving customer preferences.

Second, there is a need to explore the usage of more advanced deep learning models for improving the accuracy of sentiment classification, such as BERT, PhoBERT, or other architectures based on Transformer. These have indeed shown much better performance in understanding context and intricacies of languages, especially Vietnamese. Besides, using multi-label classification might enable the model to identify specific drivers of sentiments, such as food quality, delivery speed, pricing, and customer service.

Third, the improvement in the techniques of text preprocessing would lead to better performance of the model. This would involve the integration of NLP techniques specific to the Vietnamese language, making use of word embeddings such as Word2Vec or FastText, and developing a comprehensive Vietnamese dictionary for handling slang, abbreviations, and common misspellings. Ensemble learning methods like stacked models or hybrid models further improve classification performance. Periodic retraining of the model with the updated dataset would help a lot in maintaining performance and adaptability.

To maximize the practical benefits of sentiment analysis in the food delivery industry, businesses should integrate this technology into CRM systems. This allows companies to base strategic decisions on the customer sentiment while continuously monitoring the same and trying to improve on the quality of their services toward improving customer satisfaction. This approach will provide a significant competitive advantage in the rapidly evolving food delivery market.

REFERENCES

1. Akila, R., Revathi, S., Shreedevi, G. (2020), “Opinion Mining on Food Services using Topic Modeling and Machine Learning Algorithms”, in 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), pp. 1071-1076.

2. Akter, S., Aziz, M. T. (2016), “Sentiment analysis on facebook group using lexicon based approach”, in 2016 3rd International Conference on Electrical Engineering and Information Communication Technology (ICEEICT), pp. 1-4.

3. Balan, S., Rege, J. (2017), “Mining for social media: Usage patterns of small businesses”, Business Systems Research: The Journal of Society for Advancing Innovation and Research in Economy, Vol. 8 No. 1, pp. 43-50.

4. Burges, C. J. (1998), “A tutorial on support vector machines for pattern recognition”, Data mining and knowledge discovery, Vol. 2 No. 2, pp. 121-167.

5. Dunđer, I., Horvat, M., Lugović, S. (2016), “Word occurrences and emotions in social media: Case study on a Twitter corpus”, in Biljanović, P. (Ed.), Proceedings of the 39th International Convention on Information and Communication Technology, Electronics and Microelectronics MIPRO 2016, Croatian Society for Information and Communication Technology, Electronics and Microelectronics - MIPRO, Rijeka, pp. 1557-1560.

6. Joachims, T. (2002), Learning to classify text using support vector machines, Springer Science & Business Media.

7. Kadriu, A., Abazi, L., Abazi, H. (2019), “Albanian Text Classification: Bag of Words Model and Word Analogies”, Business Systems Research: The Journal of the Society for Advancing Innovation and Research in Economy, Vol. 10 No. 1, pp. 74-87.

CONTRIBUTION

Full name	Student's ID	Task	Contribution
Đỗ Thị Minh Phương	21070074	Model building, improvement, Crawling data	25%
Nguyễn Phương Thảo	20070981	Crawling data, Preprocessing	25%
Bế Hằng Nga	20070961	Crawling data, Report, slide	25%
Nguyễn Anh Đức	20070921	Crawling data, Visualization	25%