CHAPTER 4

RESULTS AND INITIAL FINDINGS

4.1 Introduction

This chapter aims to systematically present the core experimental process and findings of this research. It begins with the preparation and exploratory analysis of experimental data, detailing a series of data processing procedures including text cleaning and feature engineering. Subsequently, this chapter will focus on introducing how to construct and train multiple baseline and optimized models, and through empirical data, rigorously evaluate and verify the effectiveness of the multi-source information fusion strategy proposed in this study (i.e., combining BERT text features with metadata features). Finally, to deeply analyze the decision-making mechanism of the model and answer research questions (RQ2, RQ3), this chapter will introduce the SHAP interpretability analysis tool to conduct global and local attribution analysis on the best-performing model. All the findings in this chapter will provide solid data support for the final conclusion and discussion.

4.2 Dataset and Exploratory Data Analysis

The experimental data used in this study is sourced from the publicly available Yelp Open Dataset. This dataset contains a large number of user reviews for various types of businesses along with rich metadata such as user ratings (1-5 stars), business categories, geographical locations, etc. In this experiment, we selected the [specify the subset of data, for example: English reviews related to the "restaurant" category] portion, totaling [fill in the total number of samples, for example: 500,000] samples.

To gain a deeper understanding of the data characteristics, we conducted exploratory data analysis (EDA). First, we tallied the distribution of star ratings in the dataset, as shown in Figure 4.1.

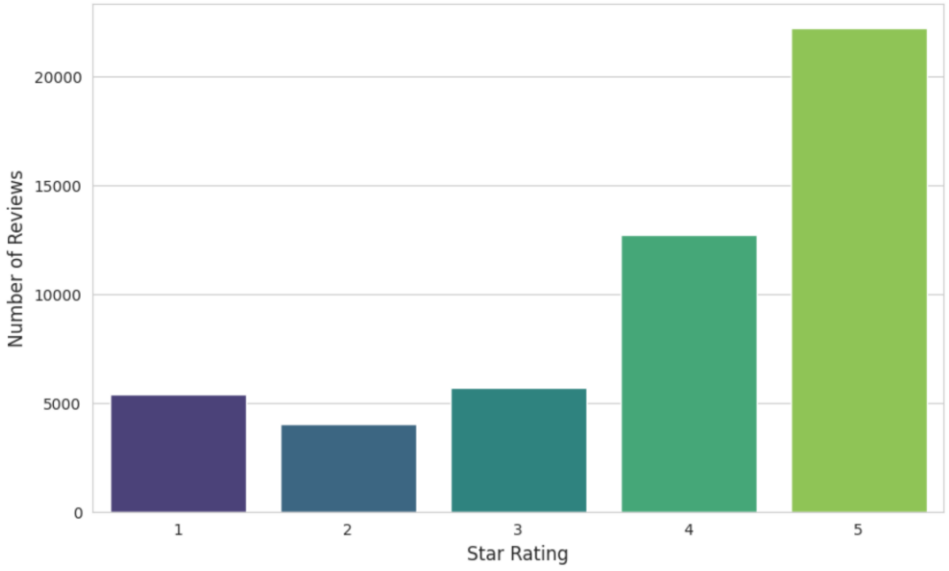

Figure 4.1 Distribution of Yelp Review Star Ratings

Figure 4.1 shows that there is a significant imbalance in the number of reviews for each star rating in the dataset. Specifically, the number of high-star (4-star and 5-star) reviews is much greater than that of low-star (1-star and 2-star) reviews, which poses a potential challenge for this study. When evaluating the model, it is particularly important to pay attention to metrics such as the Macro F1-score that are insensitive to class imbalance.

To further explore the differences in text content among different rating levels, we generated word cloud diagrams for the review texts of each star rating respectively, to visually display the high-frequency words.

Figure 4.2 Word Clouds by Star Rating

4.3 Data Preprocessing

Before feeding the data into the model for training, we carried out a series of crucial data preprocessing tasks, aiming to clean and standardize the raw data to lay a solid foundation for subsequent feature engineering and modeling.

The core of this process was the cleaning and processing of the comment text (text field). The original text data contained mixed cases and various punctuation marks (such as ! 、 ? 、 ...) And non-structured elements such as English abbreviations (like I've), etc., may all cause interference to the model in extracting effective information. Our processing operations mainly include: converting all English letters to lowercase uniformly, removing punctuation marks, and handling common English abbreviation forms.

To visually demonstrate the effect of text cleaning, Table 4.1 and 4.2 list the status comparison of some data records before and after processing.

Table 4.1 Data before Cleaning

text	stars_review	main_category	city
If you decide to eat here, just be aware it is...	3	Restaurants	North Wales
I've taken a lot of spin classes over the years...	5	Active Life	Philadelphia
Family diner. Had the buffet. Eclectic assortm...	3	Restaurants	Tucson
Wow! Yummy, different, delicious. Our favo...	5	Halal	Philadelphia
Cute interior and owner (?) gave us tour of up...	4	Sandwiches	New Orleans

Table 4.2 Data after Cleaning

text_clean	stars_review	main_category	city
if you decide to eat here just be aware it is...	3	Restaurants	North Wales
ive taken a lot of spin classes over the years...	5	Active Life	Philadelphia
family diner had the buffet eclectic assortment...	3	Restaurants	Tucson
wow yummy different delicious our favorite is...	5	Halal	Philadelphia
cute interior and owner gave us tour of upcoming...	4	Sandwiches	New Orleans

It can be clearly seen from the table that after processing, the original text has been transformed into a "cleaner" text_clean field with a uniform format, which is conducive to the subsequent learning of word vector representations by the model.

On the basis of completing the text processing, we also conducted integrity checks and duplicate removal operations on the entire dataset [Here, specific information can be supplemented, such as: A total of XX duplicate records were removed]. Ultimately, we obtained a structured dataset that can be used by the model, fully preparing for the next stage of feature engineering.

4.4 Training and Testing Split

To objectively evaluate the model's generalization ability, we divided the 49,987 data records processed in Section 4.3 into a training set and a test set. We strictly adhered to an 80:20

ratio and employed a stratified sampling strategy to ensure that the proportion of each star rating in the two datasets after division was consistent with the original data.

Ultimately, we obtained a training set containing 39,989 records and an independent test set with 9,998 records. The results show that the distribution of star ratings in the training set and the test set is almost exactly the same (for example, the proportion of 5-star reviews in the training set and the test set is 44.44% and 44.43% respectively), indicating that the stratified sampling was highly successful and laid a solid foundation for the fair comparison and reliable evaluation of subsequent models.

4.5 Feature Engineering and Fusion

4.5.1 Text Feature Extraction using BERT

The text is the core information source of this study. To capture the deep contextual and semantic information in the comments, we applied the pre-trained BERT (Bidirectional Encoder Representations from Transformers) model. Specifically, we used the [fill in the BERT model you used, for example: bert-base-uncased] model as the text encoder. For each Yelp comment, we input it into the BERT model and extract the output vector corresponding to the special start token [CLS] in the last layer. This 768-dimensional dense vector is regarded as the semantic representation of the comment text and serves as the text feature for subsequent models.

4.5.2 Metadata Feature Extraction

In addition to the text information, we also extracted structured metadata as auxiliary features. In this study, we selected [list the metadata fields you used here, such as: 'useful' vote count, business category, city of location, etc.]. As described in Section 4.3, these metadata were

transformed into numerical feature vectors after being subjected to one-hot encoding or numerical scaling.

4.5.3 Feature Fusion

To enable the model to simultaneously utilize the semantic information of the text and the structured information of the metadata, we adopted a feature fusion strategy of vector concatenation. Specifically, we concatenated the 768-dimensional text vector generated by BERT with the [fill in the dimension of the metadata feature] -dimensional metadata vector along the dimension, ultimately forming a more comprehensive fusion feature vector with a dimension of [fill in the total dimension after fusion]. This fusion vector will serve as the core of our multi-source information fusion strategy and be used for the subsequent training and evaluation of the model.

4.6 Model Training (with Default Parameters)

This step aims to establish a performance baseline for subsequent optimization and comparison. We selected three widely used classification models: Logistic Regression, Random Forest, and XGBoost. To ensure a fair comparison, we trained and evaluated these three models on two different feature sets respectively:

1. Baseline features (text only): Only using the 768-dimensional text vectors generated by BERT in Section 4.5.1.

2. Fusion features: Using the fusion feature vectors composed of text and metadata as described in Section 4.5.3.

At this stage, all models were trained with their default parameters in the scikit-learn or xgboost libraries.

4.7 Hyperparameter Tuning and Training with Best Parameters

To fully exploit the potential of each model and ensure the fairness of the comparison, we conducted systematic hyperparameter tuning for all models. We employed a combination of Grid Search and 5-fold Cross-Validation to find the optimal hyperparameter settings for each model (under both "text-only features" and "fusion features" input scenarios) on the training set. The evaluation metric for the grid search was [specify your evaluation metric, e.g., Macro F1-score]. After identifying the best parameters, GridSearchCV automatically retrained a final model using the entire training dataset with this optimal parameter set. This tuned model was then used for the final performance evaluation.

4.8 Model Evaluation

4.8.1 Performance Metrics Comparison

This section will conduct a quantitative assessment and comparison of the performance of all models on an independent test set. We plan to use accuracy, macro precision, macro recall, and macro F1-score as the core evaluation metrics. All results will be summarized in tables and charts for clear presentation. Through comparative analysis, we will: 1) identify the model with the best performance across all configurations; 2) quantitatively demonstrate the performance gains brought by the integration of metadata. Based on these results, we will determine the best performing model for this study, which will be used for subsequent in-depth analysis.

4.8.2 Misclassification Analysis of the Best Model

To delve deeper into the categories where the best model performs poorly, we plan to draw a heatmap of its confusion matrix on the test set. By observing the elements off the diagonal, we can identify the main error patterns of the model, such as which star ratings are most likely to be confused. Analyzing these error cases will help us understand the limitations of the model.

4.8.3 Explainability Analysis (SHAP) of the Best Model

To answer research questions RQ2 (Which features are the key influencing factors?) and RQ3 (What is the mechanism behind misclassification?), we plan to use the SHAP (SHapley Additive exPlanations) tool to explain the decision-making process of the best model.

- Global Feature Importance: We will generate SHAP summary plots to display the most important features for the model's global predictions, thereby identifying whether the key influencing factors come from the text or metadata.
- Local Explanation for Misclassified Cases: We will select typical misclassified cases and use SHAP force plots or waterfall plots to explain why the model made incorrect judgments, thereby revealing the internal logic and potential flaws in the model's decision-making.

4.9 Summary

This chapter will meticulously document the complete experimental process from data preparation, feature engineering to model training, evaluation and in-depth analysis. The core objective is to verify the effectiveness of the multi-source information fusion strategy and gain a deeper understanding of the model's behavior through interpretability analysis. The experimental findings of this chapter will provide solid empirical support for the final research conclusion and lay the foundation for the discussion and outlook in the next chapter.