

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

This chapter systematically elucidates the methodological design and implementation process of this research in the Yelp five-category sentiment analysis task. Addressing typical challenges in large-scale social media review data such as text ambiguity, label bias, multi-source information fusion, and model interpretability, this study aims to "improve classification accuracy, enhance generalization ability, and achieve interpretability of the decision-making process" as its core objectives, establishing a multi-stage, full-process research framework. This framework encompasses data collection and description, data preprocessing, feature engineering and fusion, model construction, performance evaluation, interpretability, and misclassification analysis. Each stage integrates previous literature with current mainstream AI technologies, emphasizing the combination of theoretical innovation and engineering practice. Through a formalized and modular design, the systematicness and scalability of the method are ensured, providing a solid technical foundation for subsequent experiments and results analysis. The following sections will detail the implementation schemes, process specifics, and theoretical basis for each stage.

3.2 Research Framework

To systematically address the multiple challenges in Yelp five-category sentiment analysis, this study designed a multi-stage methodological framework as shown in Figure 3.1. This framework not only embodies the core objectives proposed in Chapter 1 (see 1.5

Objectives) but also closely integrates the theoretical and practical requirements of the three major directions: "accuracy, generalization ability, and interpretability." The overall process is divided into seven stages:

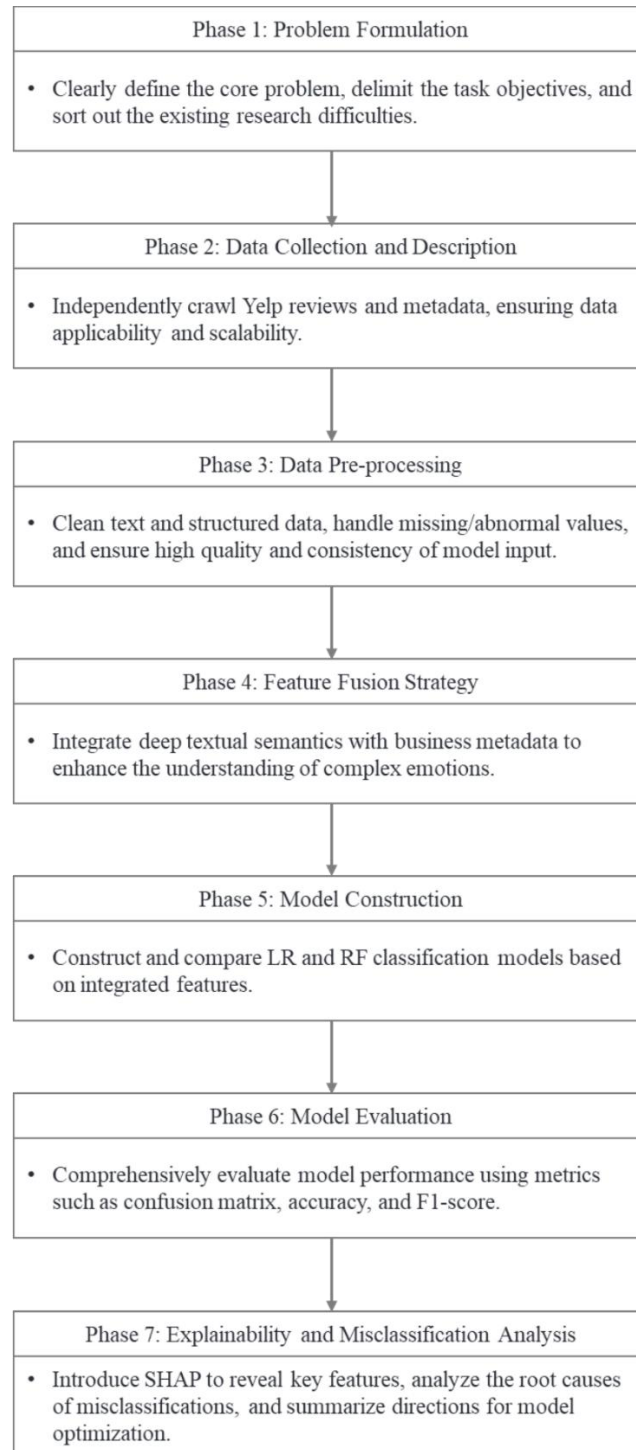


Figure 3.1 Framework diagram of research methodology workflow

The first stage focuses on key scientific challenges in sentiment analysis, such as semantic-label bias, text ambiguity, difficulties in integrating structured information, and insufficient model interpretability. This lays the theoretical foundation for introducing multi-source feature modeling and explanatory methods, directly addressing Objective 1 and Objective 2: "improving sentiment prediction accuracy and model generalization ability" and "achieving effective fusion of multi-source features."

The second stage primarily revolves around the collection and description of Yelp review texts and metadata. To address issues like the excessive volume and missing fields in public datasets, an independent crawling strategy is employed to acquire multi-dimensional fields such as review text, ratings, categories, and geographical locations. The sampling range is determined based on practical needs. This stage provides a solid foundation for subsequent data processing and feature engineering, ensuring data representativeness and diversity to support high-quality model training.

The third stage involves systematic preprocessing of text and structured metadata, including denoising, tokenization, spell correction, stop word removal, as well as missing value imputation, categorical encoding, outlier detection, and normalization. After preprocessing, the data is divided into training, validation, and test sets according to task requirements, ensuring the quality and standardization of input data, and laying the groundwork for model training and evaluation.

The fourth stage focuses on the efficient fusion of deep semantic features and business metadata. The BERT model is used for semantic encoding of review texts, and combined with metadata such as categories and geographical information, multi-source feature integration is achieved through feature concatenation or attention mechanisms, enhancing the discriminatory power of complex emotional expressions (Objective 2).

The fifth stage builds and compares various classification models based on the aforementioned features, primarily selecting Logistic Regression (LR) and Random

Forest (RF) as baseline models. This stage aims to improve model accuracy and robustness, closely aligning with Objective 1, considering data scale and engineering feasibility.

The sixth stage employs diverse evaluation metrics (accuracy, F1-score, confusion matrix, etc.) and visualization methods for a comprehensive assessment of different modeling schemes. The results provide data support for interpretability and misclassification analysis, reinforcing Objective 1 and Objective 3's requirements for overall model capability and usability.

The seventh stage introduces interpretability methods such as SHAP to conduct global and local feature contribution analysis of the model. Combined with the confusion matrix, common misclassified samples are thoroughly analyzed, revealing the model's decision-making logic and influencing factors, thereby providing theoretical basis for model optimization and practical application, echoing Objective 3.

In summary, the research methodology framework proposed in this chapter is a systematic decomposition of the research objectives in Chapter 1 and serves as the technical backbone of the entire paper. Subsequent sections will elaborate on the implementation schemes and innovations of each of the seven stages described above.

3.3 Phase 1: Problem Formulation

This stage focuses on defining the core scientific challenges in the Yelp five-category sentiment analysis task, laying the theoretical groundwork for subsequent multi-source fusion and interpretability methods. First, Yelp reviews exhibit phenomena like "semantic-label bias" and ambiguous expressions, leading to incomplete consistency between text content and ratings, which increases the complexity of sentiment modeling. Second, single textual features often fail to fully leverage Yelp's rich structured metadata (e.g., categories, geographical location), limiting the model's ability to understand context.

Furthermore, while deep learning models possess powerful representation capabilities, they lack interpretability, making it difficult to meet practical business demands for decision transparency. Concurrently, issues such as imbalanced label distribution and easy confusion between adjacent star ratings in the five-category task also increase the difficulty of training and evaluation.

In light of these challenges, this study defines the tasks for this stage as:

1. Scientific Problem and Objective Definition: Systematize the typical challenges in Yelp sentiment analysis and clearly define the objectives of improving accuracy, generalization ability, and interpretability.
2. Task Refinement: Focus on key issues such as text and metadata fusion, model interpretability, and misclassification traceback, providing guidance for method design.
3. Phase-specific Output: Lay the theoretical and practical foundation for subsequent stages, including data collection, feature engineering, model construction, and interpretability analysis.

The theoretical review and problem definition in this stage will serve as the starting point for the subsequent research process and experimental design, ensuring that the entire workflow efficiently revolves around practical business needs and scientific objectives.

3.4 Phase 2: Data Collection and Description

3.4.1 Data Acquisition Method

During the data acquisition phase, this study first examined the official Yelp Academic Dataset. As shown in Figure 3.2, the raw dataset comprises multiple large JSON files containing reviews, businesses, users, etc. (e.g., the `yelp_academic_dataset_review.json` file alone exceeds 5GB, as seen in Figure 3.2). The

overall uncompressed data volume is extremely large, with fields widely dispersed, making processing and uploading very difficult. Furthermore, critical information such as business categories and city is missing for some reviews in the official dataset, and fields require complex ID associations, increasing the technical difficulty of data preprocessing. These practical issues have been repeatedly mentioned in prior literature (Taboada et al., 2011; Rodríguez-Ibáñez et al., 2023).










Name	Date modified	Type	Size
 Dataset_User_Agreement.pdf	2/16/2022 6:03 AM	PDF File	79 KB
 Yelp Dataset Documentation & ToS copy.pdf	1/8/2025 3:55 AM	PDF File	122 KB
 yelp_academic_dataset_business.json	1/20/2022 6:35 AM	JSON File	116,078 KB
 yelp_academic_dataset_checkin.json	1/20/2022 6:39 AM	JSON File	280,234 KB
 yelp_academic_dataset_review.json	1/20/2022 6:51 AM	JSON File	5,216,669 KB
 yelp_academic_dataset_tip.json	1/20/2022 6:40 AM	JSON File	176,372 KB
 yelp_academic_dataset_user.json	1/20/2022 6:39 AM	JSON File	3,284,501 KB
 yelp_dataset.tar	1/8/2025 12:39 AM	Compressed File (TAR)	4,242,083 KB
 yelp_dataset-2.tar	2/16/2022 6:04 AM	Compressed File (TAR)	9,073,940 KB

Figure 3.2 Yelp Open Dataset

Therefore, to ensure complete sample fields, flexible sampling structure, and efficient experimentation, this study opted to develop its own web crawler program for targeted collection of Yelp reviews and their structured metadata. The specific process is illustrated in Figure 3.3: based on business requirements, the crawler is configured to collect data within specified cities, business categories, and review timeframes, automatically acquiring sample data with multi-dimensional information such as review text, star ratings, categories, and city. The data collection process adheres to the Yelp platform's robots.txt protocol, ensuring the legality and compliance of the data source.

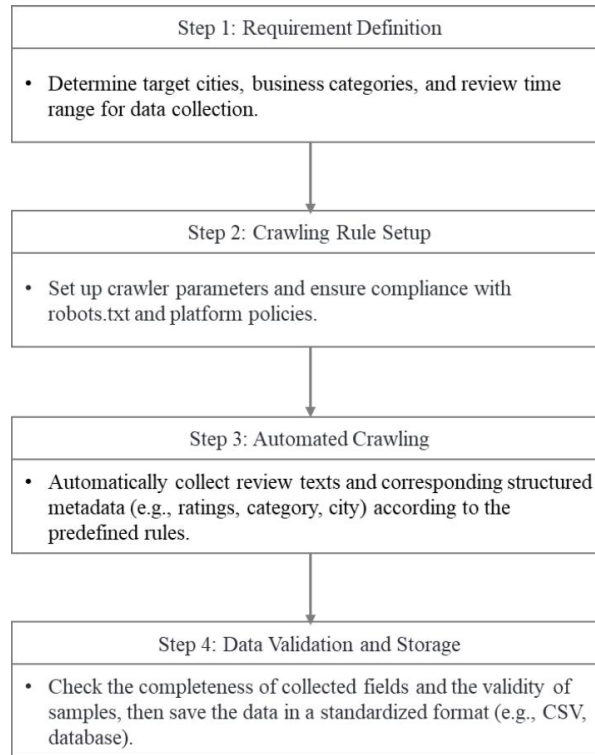


Figure 3.3 Data collection flowchart

3.4.2 Field Selection and Sample Scope

3.4.2.1 Field Selection

Given mainstream modeling practices and literature experience in sentiment analysis tasks (Taboada et al., 2011), this study primarily collected and utilized the following fields:

- Review text: The core information source for sentiment classification.
- Star rating: A five-level label serving as the target variable for supervised learning tasks.
- Business category: Used for feature fusion and business analysis to enhance the model's generalization ability.

- City: Introduced to account for geographical diversity and analyze the relationship between sentiment and region.

3.4.2.2 Sample Scope

- Select only English reviews to avoid multilingual interference.
- Cover multiple cities and categories to ensure sample diversity and generalization ability.
- The total sample size will be determined based on experimental feasibility and model complexity, typically ranging from tens of thousands to over a hundred thousand entries, with the exact number to be finalized during subsequent collection.

3.4.3 Data Basic Description and Distribution Analysis

After completing the data collection, the samples were first subjected to descriptive statistics, including:

- Star rating distribution (1–5 stars), to check for data balance and implement downsampling or oversampling measures if necessary.
- Sample proportion across different categories and cities, to identify any extreme imbalances or data anomalies.
- Proportion and distribution of missing field values.

3.5 Phase 3: Data Pre-processing

This stage primarily involves systematic cleaning and standardization of the raw Yelp review data and its structured metadata to ensure data quality and consistency for subsequent feature engineering and model training.

3.5.1 Text Data Cleaning

- First, the review text undergoes noise cleaning, including the removal of HTML tags, special characters, emojis, and superfluous spaces.
- Natural Language Processing (NLP) tools (e.g., NLTK, spaCy) are used for tokenization, segmenting long texts into word or subword sequences.
- All text is uniformly converted to lowercase to reduce vocabulary sparsity.
- Spelling correction algorithms are applied to rectify common misspellings, improving the standardization of model input.
- Stop words (e.g., "the," "is," "and" – common words with no actual semantic meaning) are removed, retaining only meaningful information.
- In consideration of sentiment analysis and BERT feature requirements, texts undergo lemmatization or stemming to normalize different word forms, which facilitates semantic understanding and feature extraction.

3.5.2 Structured Data Handling

- For structured metadata (such as business category, city), check for missing values. For missing information, options include excluding samples or imputing missing fields, with the specific strategy determined by data distribution.
- Standardize categorical fields (e.g., category, city) to address issues like synonyms and spelling variations, ensuring field consistency.
- Encode categorical fields into numerical variables, commonly using One-Hot Encoding or Label Encoding, in preparation for subsequent feature integration.
- Check numerical fields like ratings for outliers or invalid values, and perform appropriate corrections or filtering.
- For continuous metadata (e.g., price range, review length), normalize or standardize as needed.

3.5.3 Dataset Splitting

After completing data cleaning and standardization, samples are randomly divided into training, validation, and test sets according to a certain proportion (e.g., 8:1:1), ensuring that the distribution of star ratings across different subsets is as consistent as possible (i.e., stratified sampling).

Each subset, after division, requires statistical description (e.g., category distribution, text length distribution) to ensure the fairness and scientific rigor of the experimental evaluation.

3.5.4 Quality Assessment and Visualization

Upon completion of preprocessing, the overall data quality is assessed by analyzing metrics such as the proportion of remaining missing values and the balance of category distribution; any issues discovered should be promptly addressed.

Visualization tools (e.g., bar charts, pie charts) are utilized to display data distribution, facilitating an intuitive understanding of the data structure, with relevant figures cited in the main text.

3.6 Phase 4: Feature Fusion Strategy

This phase primarily focuses on how to effectively integrate text semantic features with structured metadata (e.g., business category, city) to enhance the sentiment analysis model's ability to represent and discriminate complex information.

3.6.1 Textual Feature Extraction with BERT

In this study, the extraction of text semantic features primarily relies on the pre-trained BERT (Bidirectional Encoder Representations from Transformers) model. BERT

possesses strong contextual understanding capabilities, enabling it to capture complex semantic relationships, contextual information, and long-range dependencies within review texts. Specifically, after tokenizing and standardizing the raw review text, it is input into the BERT model for encoding. Through its multi-layer bidirectional Transformer structure, BERT maps each review into a high-dimensional dense semantic vector, fully preserving the nuanced differences in emotional expression and latent semantic features within the text. These vectors will serve as the main text input for subsequent feature fusion and sentiment classification models, effectively enhancing the model's ability to discriminate between different emotions and expression styles.

3.6.2 Metadata Feature Construction

In addition to text features, structured metadata such as business category, city, and review timestamp also play a significant supplementary role in sentiment analysis results. To fully leverage this non-textual information, this study standardizes and digitizes all metadata fields. Specific practices include: converting discrete variables like category and city into vector forms using One-Hot Encoding or Label Encoding to suit subsequent feature concatenation and modeling requirements. Review timestamps can be further used to extract auxiliary features such as "is_weekend" or "season" to enhance the model's sensitivity to temporal or regional differences. The systematic processing of the aforementioned metadata enables the model to consider multi-dimensional information from text, context, and objects, achieving sentiment classification with greater generalization ability and business interpretability.

3.6.3 Fusion Method

To fully leverage the complementary advantages of multi-source features, this study explored various feature fusion strategies. The most straightforward approach is vector concatenation, where the text semantic vectors generated by BERT are directly concatenated dimensionally with the encoded metadata features, forming a unified high-dimensional input feature. This method is simple to implement and easily integrated into

mainstream machine learning frameworks. Additionally, to address the issue of dynamically changing weights for multi-source information, this study will also investigate attention mechanism fusion. By introducing an attention layer, the model can automatically learn the contribution weights of different features to the final classification result, thereby more intelligently fusing text and metadata information. Through the comparison of these multi-strategies, the research can systematically evaluate the impact of different fusion methods on model accuracy, generalization ability, and interpretability, providing theoretical support for subsequent interpretability analysis and business decisions.

3.7 Phase 5: Model Construction

This phase aims to systematically construct an efficient and interpretable multi-class sentiment analysis model for Yelp reviews, based on the aforementioned feature fusion results. The model structure, selection criteria, and process design are as follows:

3.7.1 Model Selection and Rationale

This study primarily employs Logistic Regression (LR) and Random Forest (RF) as the main classification models, with inputs consisting of BERT deep semantic features and structured metadata. The rationale for model selection is as follows:

- **Logistic Regression (LR)**

LR models possess excellent multi-class classification capabilities, particularly suitable for high-dimensional sparse feature data (such as TF-IDF, BERT vectors), and offer good engineering efficiency and interpretability. Literature indicates (Sharma et al., 2024; Taboada et al., 2011) that LR performs stably in large-scale sentiment classification tasks and is easily amenable to parameter tuning for fused features. Therefore, LR is selected as a robust baseline model.

- **Random Forest (RF)**

RF models can automatically capture complex non-linear relationships between features, thereby improving classification accuracy and generalization ability in scenarios where text and metadata are fused. Additionally, RF supports quantitative analysis of feature importance, which facilitates the interpretation of model decision logic. Drawing from the experiences of Rodríguez-Ibáñez et al. (2023) and others, RF has become a mainstream choice for multi-source fusion and complex-structured sentiment classification tasks.

3.7.2 Model Architecture and Implementation Strategy

The model architecture in this study adopts a hierarchical design to ensure that multi-source fused features can be efficiently passed to downstream primary classifiers, and to facilitate subsequent expansion and interpretability analysis. The overall architecture is shown in Figure 3.4.

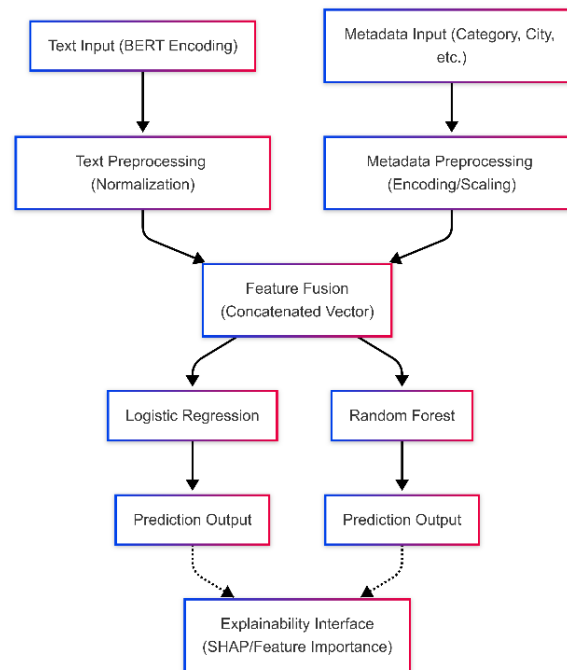


Figure 3.4 The Vertical Architecture of Multi-source Sentiment Classification Model

3.7.3 Training and Validation Workflow

To ensure the model possesses good generalization ability and fair comparability, this study systematically designed the model training and validation process, which includes the following key steps:

1. Dataset Splitting

The cleaned and feature-fused dataset is divided into training, validation, and test sets in an 8:1:1 ratio. Stratified sampling is used for splitting to ensure consistent class distribution across subsets.

2. Cross-Validation and Parameter Tuning

For both LR and RF models, K-fold cross-validation (e.g., 5-fold) and grid search are employed for hyperparameter optimization. For LR, the regularization coefficient is primarily adjusted, while for RF, parameters such as the number of decision trees and maximum depth are optimized.

3. Model Training and Evaluation

Models are trained on the training set, and metrics (accuracy, F1-score, etc.) are monitored on the validation set to select the optimal parameter scheme.

4. Performance Testing and Archiving

The final model is evaluated on the test set, outputting multiple metrics (Accuracy, Precision, Recall, F1-score, Confusion Matrix, etc.). All model parameters, procedures, and evaluation results are fully archived in a Jupyter Notebook for easy reproduction and tracking.

3.7.4 Consideration and Exclusion of Other Algorithms

During the model design process, this study systematically evaluated the applicability of algorithms such as SVM and single decision trees but ultimately did not include them in the main experimental group. The specific reasons are as follows:

- **SVM (Support Vector Machine)**

In environments with high-volume, high-dimensional sparse features, SVM models incur significant training and inference times, and resource consumption is considerably higher than that of LR (Logistic Regression) and RF (Random Forest) models. Based on literature review and engineering efficiency, this project prioritized mainstream models that offer efficient scalability and ease of hyperparameter tuning.

- **Single Decision Tree**

While single decision trees offer excellent interpretability, their accuracy in high-dimensional, multi-class tasks is limited, and they are highly prone to overfitting, performing significantly worse than ensemble models (like Random Forest). Therefore, they were only considered as a tool for auxiliary interpretability analysis, not as a primary classification model.

- **Deep Neural Networks**

At this stage, the focus was on engineering efficiency, interpretability, and experimental reproducibility. Consequently, deep neural network models, which require extensive hyperparameter tuning and substantial computational resources, were not introduced.

3.8 Phase 6: Model Evaluation

This stage primarily focuses on model performance evaluation and result visualization, systematically examining the effectiveness of different modeling

approaches to provide a basis for subsequent interpretability analysis and model optimization. The evaluation process includes an explanation of the metric system, visualization schemes, comparative experiments, and statistical analysis, with specific content as follows:

3.8.1 Evaluation Metrics

To comprehensively evaluate the performance of the multi-class sentiment analysis model, this study employs the following mainstream evaluation metrics:

- **Accuracy:** Measures the proportion of correctly classified predictions among all predictions made by the model, reflecting its overall classification capability. It is suitable for scenarios with relatively balanced sample distributions, but for imbalanced classes, it needs to be analyzed in conjunction with other metrics.
- **Precision and Recall:** These metrics quantify the model's false positive and false negative rates for each class, respectively. Precision indicates the proportion of samples predicted as a certain class that are actually of that class, while recall indicates the proportion of actual samples of a certain class that are correctly predicted. These two can be further combined into the F1-score.
- **F1-score:** The harmonic mean of precision and recall, balancing both accuracy and coverage. It is suitable for comprehensive evaluation in multi-class and class-imbalanced scenarios. The specific calculation method is:

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

- **Confusion Matrix:** It is used to comprehensively display the model's classification results across various categories, including the number of correct predictions and misclassifications between different classes. By analyzing the

confusion matrix, one can identify which categories the model tends to confuse, providing a basis for subsequent misclassification analysis.

3.8.2 Visualization and Comparative Analysis Plan

To enhance the intuitiveness and interpretability of the model evaluation results, this study employs various visualization methods to display and compare experimental outcomes:

- **Bar Charts and Line Charts**

These charts present comparative results of different models across various evaluation metrics (e.g., Accuracy, F1-score), highlighting how multi-source fusion features improve model performance.

- **Confusion Matrix Heatmap**

This visualization displays the confusion matrix results in a heatmap format, making it easier to identify frequently confused categories and localized model deficiencies, thereby aiding subsequent analysis of misclassified samples.

- **Per-Category Metric Distribution Plots**

These plots analyze the distribution of Precision, Recall, and F1-score for each category. They help assess the model's ability to distinguish between minority and majority classes and evaluate whether the model exhibits class bias.

3.8.3 Comparative Experiment Design

To validate the effectiveness of the proposed multi-source feature fusion strategy, this study conducts systematic comparisons under the following experimental scenarios:

- **Baseline Model vs. Main Model Comparison**

This involves comparing the performance of models that use only text features (e.g., TF-IDF+LR, BERT+LR) as baselines against multi-source feature fusion models (BERT+Metadata+LR/RF).

- **Various Feature Combination Experiments**

Experiments are designed with three categories: "text features only," "metadata features only," and "text and metadata fusion." This allows for analysis of the performance improvement achieved by multi-source fusion across different model architectures.

- **Stability Analysis Under Different Model Parameter Settings**

The models are repeatedly trained under multiple hyperparameter configurations to examine the stability and robustness of model performance, ensuring the universality and reproducibility of the experimental conclusions.

3.9 Phase 7: Explainability and Misclassification Analysis

This stage focuses on model interpretability and misclassification sample analysis, aiming to reveal the logic behind model decisions, feature contributions, and common causes of misclassification. By employing interpretability tools such as SHAP, we provide global and local explanations of model outputs. Concurrently, by combining the confusion matrix with typical samples, we thoroughly analyze model shortcomings and areas for optimization, offering theoretical support for practical business applications and subsequent model improvements.

3.9.1 SHAP-based Explanation Method

To enhance the transparency and trustworthiness of the sentiment analysis model, this study adopts the SHAP (SHapley Additive exPlanations) method for interpretability

analysis of the multi-source feature fusion model. SHAP is a game-theoretic explanation framework capable of quantifying the global and local contribution of each feature to the model's prediction results.

- **Global Feature Importance Analysis**

By statistically analyzing the SHAP values, we determine the average contribution of different features (e.g., BERT semantic vectors, merchant categories, cities) across all samples. This helps identify which features are most influential in the model's overall judgment. The specific analysis results will be presented through visualizations such as bar charts in the experimental analysis section.

- **Local Explanation and Case Study**

For individual samples, SHAP can explain why a particular sentiment category was predicted, specifically by showing the positive and negative contributions of each feature to the current prediction. This can be visualized using waterfall plots, force plots, or similar methods, and detailed analysis will be presented in the experimental section.

3.9.2 Misclassification Analysis Scheme

To further optimize model performance and enhance practical application value, this study systematically analyzes common misclassification phenomena and their causes in the five-class task by combining the confusion matrix with typical cases.

- **Identification of Confused Categories**

Based on the confusion matrix heatmap, we identify the star rating pairs that the model most frequently confuses, quantifying the misclassification proportion between different categories. Specific visualizations will be provided in the subsequent experimental section.

- **Attribution of Typical Misclassified Samples**

Representative samples with high-frequency misclassifications are selected, and combined with SHAP local explanations, we analyze the main features leading to the model's incorrect classification. Relevant visual analyses will be provided in the experimental results section.

- **Optimization Recommendations for Business Scenarios**

For categories where the model frequently gets confused, we propose optimization directions—such as introducing more auxiliary features or adjusting class weights—in conjunction with actual business requirements and data distribution.

3.10 Summary

Chapter 3 has presented the research methodology that provides a structured approach to addressing the key research objectives of this thesis. The proposed methodological framework is composed of seven main phases, including problem formulation, data collection and description, data pre-processing, feature fusion strategy, model construction, model evaluation, and explainability with misclassification analysis. Specifically, the workflow integrates advanced techniques such as BERT-based semantic feature extraction, multi-source metadata fusion, and SHAP-based model interpretability. Each phase is designed to ensure the effectiveness, robustness, and transparency of the overall sentiment classification system for Yelp reviews. The research methodology outlined in this chapter serves as a comprehensive guideline for the experimental procedures and subsequent analysis. The next chapter will present the experimental results and performance evaluation based on the methodology established here.

REFERENCES

- Rodríguez-Ibáñez, M., Casáñez-Ventura, A., Castejón-Mateos, F., & Cuenca-Jiménez, P. M. (2023). A review on sentiment analysis from social media platforms. *Expert Systems with Applications*, 223, 119862.
- Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2), 267-307.