

AN INTERPRETABLE BERT-BASED SENTIMENT CLASSIFICATION WITH
METADATA FUSION FOR YELP REVIEWS

GAO JINGKAI

UNIVERSITI TEKNOLOGI MALAYSIA



UNIVERSITI TEKNOLOGI MALAYSIA
DECLARATION OF Choose an item.

Author's full name : GAO JINGKAI

Student's Matric No. : MCS241032 Academic Session : 2024-25/02

Date of Birth : 09/14/1999 UTM Email : gaojingkai@graduate.utm.my

Choose an item. Title : AN INTERPRETABLE BERT-BASED SENTIMENT CLASSIFICATION WITH METADATA FUSION FOR YELP REVIEWS

I declare that this AN INTERPRETABLE BERT-BASED SENTIMENT CLASSIFICATION WITH METADATA FUSION FOR YELP REVIEWS is classified as:

☒

OPEN ACCESS

I agree that my report to be published as a hard copy or made available through online open access.

☐

RESTRICTED

Contains restricted information as specified by the organization/institution where research was done.
(The library will block access for up to three (3) years)

☐

CONFIDENTIAL

Contains confidential information as specified in the Official Secret Act 1972)

(If none of the options are selected, the first option will be chosen by default)

I acknowledged the intellectual property in the Choose an item. belongs to Universiti Teknologi Malaysia, and I agree to allow this to be placed in the library under the following terms :

1. This is the property of Universiti Teknologi Malaysia
2. The Library of Universiti Teknologi Malaysia has the right to make copies for the purpose of research only.
3. The Library of Universiti Teknologi Malaysia is allowed to make copies of this Choose an item. for academic exchange.

Signature of Student: GAO JINGKAI

Signature : GAO JINGKAI

Full Name: GAO JINGKAI

Date : 06/30/2025

Approved by Supervisor(s)

Signature of Supervisor I:

Signature of Supervisor II

Full Name of Supervisor I
 NOOR HAZARINA HASHIM

Full Name of Supervisor II
 MOHD ZULI JAAFAR

Date :

Date :

NOTES : If the thesis is CONFIDENTIAL or RESTRICTED, please attach with the letter from the organization with period and reasons for confidentiality or restriction

“Choose an item. hereby declare that Choose an item. have read this Choose an item.
and in Choose an item.
opinion this Choose an item. is sufficient in term of scope and quality for the
award of the degree of Choose an item.”

Signature : _____
Name of Supervisor I : KHAIRUR RIJAL JAMALUDIN
Date : 9 MAY 2017

Signature : _____
Name of Supervisor II : NOOR HAZARINA HASHIM
Date : 9 MAY 2017

Signature : _____
Name of Supervisor III : MOHD ZULI JAAFAR
Date : 9 MAY 2017

Choose an item.Choose an item.

School of Education
Faculty of Social Sciences and Humanities
Universiti Teknologi Malaysia

ACKNOWLEDGEMENT

In preparing this thesis, I was in contact with many people, researchers, academicians, and practitioners. They have contributed towards my understanding and thoughts. In particular, I wish to express my sincere appreciation to my main thesis supervisor, Professor Dr. Mohd Shariff Nabi Baksh, for encouragement, guidance, critics and friendship. I am also very thankful to my co-supervisor Professor Dr Awaluddin Mohd Shahrour and Associate Professor Dr. Hishamuddin Jamaluddin for their guidance, advices and motivation. Without their continued support and interest, this thesis would not have been the same as presented here.

I am also indebted to Universiti Teknologi Malaysia (UTM) for funding my Ph.D study. Librarians at UTM, Cardiff University of Wales and the National University of Singapore also deserve special thanks for their assistance in supplying the relevant literatures.

My fellow postgraduate student should also be recognised for their support. My sincere appreciation also extends to all my colleagues and others who have provided assistance at various occasions. Their views and tips are useful indeed. Unfortunately, it is not possible to list all of them in this limited space. I am grateful to all my family member.

ABSTRACT

The purpose of this study is to investigate the application of genetic algorithm (GA) in modelling linear and non-linear dynamic systems and develop an alternative model structure selection algorithm based on GA. Orthogonal least square (OLS), a gradient descent method was used as the benchmark for the proposed algorithm. A model structure selection based on modified genetic algorithm (MGA) has been proposed in this study to reduce problems of premature convergence in simple GA (SGA). The effect of different combinations of MGA operators on the performance of the developed model was studied and the effectiveness and shortcomings of MGA were highlighted. Results were compared between SGA, MGA and benchmark OLS method. It was discovered that with similar number of dynamic terms, in most cases, MGA performs better than SGA in terms of exploring potential solution and outperformed the OLS algorithm in terms of selected number of terms and predictive accuracy. In addition, the use of local search with MGA for fine-tuning the algorithm was also proposed and investigated, named as memetic algorithm (MA). Simulation results demonstrated that in most cases, MA is able to produce an adequate and parsimonious model that can satisfy the model validation tests with significant advantages over OLS, SGA and MGA methods. Furthermore, the case studies on identification of multivariable systems based on real experiment data from two systems namely a turbo alternator and a continuous stirred tank reactor showed that the proposed algorithm could be used as an alternative to adequately identify adequate and parsimonious models for those systems. Abstract must be bilingual. For a thesis written in Bahasa Melayu, the abstract must first be written in Bahasa Melayu and followed by the English translation. If the thesis is written in English, the abstract must be written in English and followed by the translation in Bahasa Melayu. The abstract should be brief, written in one paragraph and not exceed one (1) page. An abstract is different from synopsis or summary of a thesis. It should states the field of study, problem definition, methodology adopted, research process, results obtained and conclusion of the research.

ABSTRAK

Kajian ini dilakukan bertujuan mengkaji penggunaan algoritma genetik (GA) dalam pemodelan sistem dinamik linear dan tak linear dan membangunkan kaedah alternatif bagi pemilihan struktur model menggunakan GA. Algoritma kuasa dua terkecil ortogon (OLS), satu kaedah penurunan kecerunan digunakan sebagai bandingan bagi kaedah yang dicadangkan. Pemilihan struktur model menggunakan kaedah algoritma genetik yang diubahsuai (MGA) dicadangkan dalam kajian ini bagi mengurangkan masalah konvergensi pramatang dalam algoritma genetik mudah (SGA). Kesan penggunaan gabungan operator MGA yang berbeza ke atas prestasi model yang terbentuk dikaji dan keberkesanan serta kekurangan MGA ditandakan. Kajian simulasi dilakukan untuk membandingkan SGA, MGA dan OLS. Dengan menggunakan bilangan parameter dinamik yang setara kajian ini mendapati, dalam kebanyakan kes, prestasi MGA adalah lebih baik daripada SGA dalam mencari penyelesaian yang berpotensi dan lebih berkebolehan daripada OLS dalam menentukan bilangan sebutan yang dipilih dan ketepatan ramalan. Di samping itu, penggunaan variasi tempatan dalam MGA untuk menambah baik algoritma tersebut dicadangkan dan dikaji, dinamai sebagai algoritma memetik (MA). Hasil simulasi menunjukkan, dalam kebanyakan kes, MA berkeupayaan menghasilkan model yang bersesuaian dan parsimoni dan memenuhi ujian pengesahan model di samping memperoleh beberapa kelebihan dibandingkan dengan kaedah OLS, SGA dan MGA. Tambahan pula, kajian kes untuk sistem berbilang pemboleh ubah menggunakan data eksperimental sebenar daripada dua sistem iaitu sistem pengulang-alik turbo dan reaktor teraduk berterusan menunjukkan algoritma ini boleh digunakan sebagai alternatif untuk memperoleh model termudah yang memadai bagi sistem tersebut.

TABLE OF CONTENTS

TITLE

PAGE

ACKNOWLEDGEMENT	iii
ABSTRACT	iv
ABSTRAK	v
TABLE OF CONTENTS	vi
 CHAPTER 1 INTRODUCTION	 1
1.1 Introduction	1
1.2 Problem Background	2
1.3 Statement of the Problem	4
1.4 Research Questions	5
1.5 Research Aim and Objectives	5
1.6 Scope of the Study	6
1.7 Significance of the Research	7
1.8 Structure of the Thesis	8
1.9 Summary	9
 CHAPTER 2 LITERATURE REVIEW	 11
2.1 Introduction	11
2.2 Application Background and Motivation	12
2.3 Overview and Comparison of Main Sentiment Analysis Datasets	13
2.4 Overview and Comparative Analysis of Sentiment Analysis Methods	18
2.5 Overview of Explainability Methods in Sentiment Analysis	30
2.6 Misclassification Analysis and Evaluation of Multi- class Sentiment Modeling	33
2.7 Research Gaps	35
2.8 Summary	37
 CHAPTER 3 RESEARCH METHODOLOGY	 39
3.1 Introduction	39
3.2 Research Framework	39
3.3 Phase 1: Problem Formulation	42
3.4 Phase 2: Data Collection and Description	43

3.5	Phase 3: Data Pre-processing	46
3.6	Phase 4: Feature Fusion Strategy	48
3.7	Phase 5: Model Construction	49
3.8	Phase 6: Model Evaluation	53
3.9	Phase 7: Explainability and Misclassification Analysis	55
3.10	Summary	57
CHAPTER 4	RESULTS AND INITIAL FINDINGS	59
4.1	Introduction	59
4.2	Dataset and Exploratory Data Analysis	59
4.3	Data Preprocessing	61
4.4	Training and Testing Split	63
4.5	Feature Engineering and Fusion	63
4.6	Model Training (with Default Parameters)	64
4.7	Hyperparameter Tuning and Training with Best Parameters	65
4.8	Model Evaluation	65
4.9	Summary	66
CHAPTER 5	CONCLUSION AND FUTURE WORKS	67
5.1	Introduction	67
5.2	Research Summary and Conclusion	67
5.3	Research Limitations	69
5.4	Future Works	69
	REFERENCES	71

CHAPTER 1

INTRODUCTION

1.1 Introduction

Sentiment analysis, being a fundamental component of Natural Language Processing (NLP), is extensively applied in areas such as product sentiments and tracking of public sentiments. Especially in consideration of the advances achieved through deep learning technology, sentiment analysis has come to prominence (Sharma et al., 2024). Concurrently, User-Generated Content (UGC) is also becoming increasingly important in its own right, with companies relying significantly on it to gauge consumer opinion.

Popular international review sites like Yelp have a vast library of user reviews and text opinions. As the data reflects individuals' own sentiments and offers guiding value, researchers have used it frequently for sentiment analysis (Xu et al., 2015). However, Yelp reviews often have the "semantic-label bias" where the expression of emotions does not match the ratings. The text sometimes expresses dissatisfaction but has a five-star rating. In addition, there are also problems such as ambiguous language. These features significantly enhance the difficulty in sentiment judgment, especially in fine-grained five-class rating (1 to 5 stars), where misclassification of neighboring ratings tends to occur (Xu et al., 2015).

As a result, in an effort to improve tackling such textual issues, researchers have increasingly turned to deep semantic modeling techniques. Pre-trained language models (PLMs) such as Bidirectional Encoder Representations from Transformers (BERT), due to their robust semantic comprehension abilities, have become state-of-the-art methods in sentiment analysis, showing efficiency in processing ambiguous emotional utterances (Devlin et al., 2019; Rodríguez-Ibáñez et al., 2023). Yet these models are "black boxes," i.e., it is difficult to ascertain how they are making their

decisions, which limits their application in business contexts in which high trust is required (Rogers et al., 2021).

Furthermore, most current research has the inclination to only examine textual content, ignoring large amounts of structured data present on websites such as Yelp, such as geographical location and business category. In fact, such data can provide valuable contextual cues, and its fusion with textual content can lead to more accurate sentiment detection (Rodríguez-Ibáñez et al., 2023).

Due to the problems in Yelp reviews, including unclear emotional expressions, rating-content inconsistencies, and inability to take advantage of structured information, it is the objective of this study to build a five-class sentiment prediction model that combines several sources of information as well as explanation mechanisms. This model will make decisions by taking into account both the deep semantics of text and the structured business attributes in a holistic manner, and also employ explanation methods such as SHapley Additive exPlanations (SHAP) (Lundberg & Lee, 2017) to break down the decision-making process of the model and check for any biases. Finally, the goal is not only to improve classification accuracy but also to render the model's prediction process more explainable, interpretable, and trustworthy and thus offer more meaningful data hints for platform service assessment and business optimization.

1.2 Problem Background

Now, with the advent of UGC on an unprecedented scale, applications of sentiment analysis in domains such as social media sentiment and product reviews have grown progressively vital (Sharma et al., 2024). Websites such as Yelp, by virtue of their combination of structured ratings and users' subjective written reviews, have emerged as a popular data source for sentiment modeling (Xu et al., 2015). Despite this, this research faces some challenges in conducting sentiment classification on Yelp reviews.

Firstly, there is the presence of "semantic-label bias" in Yelp reviews, i.e., the rating given by the user does not align with the sentiment of their reviews (e.g., a very negative review with a high rating). Secondly, the language of reviews is generally informal and vague. These phenomena prevent models from being capable of identifying the actual sentiment of the user precisely, particularly in fine-grained classification such as 1 to 5 stars where models can easily confuse neighboring ratings (Xu et al., 2015; Ravi & Ravi, 2015; Taboada et al., 2011). Second, while pre-trained language models such as BERT have certainly achieved impressive advances in interpreting the meaning of text (Devlin et al., 2019; Rogers et al., 2021), they are inevitably "black boxes," and therefore it is not straightforward for us to understand how they are making their decisions. This renders them less appropriate for deployment in certain mission-critical business applications (Rogers et al., 2021).

Besides, many current studies tend to focus only on the text, often missing out on the valuable structured information Yelp provides, like business categories and locations. This kind of data can offer really important context, and it's generally thought that mixing it with textual analysis will help us get a more accurate picture of what users are trying to say (Rodríguez-Ibáñez et al., 2023; Zhao et al., 2021). Synchronously, while explainability methods like SHAP (Lundberg & Lee, 2017) can reveal the decision basis of complex models to enhance model transparency and understand prediction biases (e.g., confusion between 4 and 5 stars), their systematic application in multi-class sentiment modeling on Yelp remains insufficient (Qu et al., 2022).

To sum up, classifying sentiment in Yelp reviews comes with a few key hurdles: understanding the nuances of the text can be tricky, many popular models lack transparency, and this research are often not making full use of all the different information sources available. This study sets out to build a five-class sentiment prediction framework that brings together deep textual understanding, structured business details, and SHAP explanation methods. Its objective is, on one hand, to improve classification accuracy, and on the other, to make the model's prediction process more transparent, understandable, and reliable. This, in turn, will provide

genuinely robust data insights for platform service evaluation and business decision-making.

1.3 Statement of the Problem

Even though sentiment analysis is pretty common these days for all sorts of UGC, we still run into three main hurdles when trying to build detailed sentiment models, especially for something like Yelp's five-star rating system. These problems really hold back how useful current models are in practice and how much more we could learn from them.

First of all, many existing models have a tough time with the tricky meanings in language. A big one is "semantic-label bias" (Xu et al., 2015), which is basically when a user's star rating doesn't quite match what they're saying in their review. This really hurts how well they can classify tricky cases, like telling a 4-star review from a 5-star one (Ravi & Ravi, 2015). On top of that, these models often struggle to make sense of language that's ambiguous or unclear (Taboada et al., 2011).

Secondly, it's just plain hard to understand how these advanced models make their decisions. Sure, pre-trained models like BERT have gotten better at understanding what words mean (Devlin et al., 2019), but they often work like "black boxes." This makes it tough to see their decision-making process, which is a problem when you need to really trust the results or want to check for biases (Rogers et al., 2021; Lundberg & Lee, 2017).

And lastly, we're often not making the most of structured information. Most current studies tend to ignore valuable structured data on Yelp, like business categories. This means they're not effectively mixing different types of information to get a clearer picture of what users are really trying to say (Rodríguez-Ibáñez et al., 2023; Zhao et al., 2021).

Consequently, there is a lack of a unified modeling framework in the current sentiment analysis field that can systematically integrate textual semantics, structured features, and provide reliable model explanations. This constitutes the core gap that this study aims to bridge.

1.4 Research Questions

This study will focus on the following three core questions:

RQ1: How to build a Yelp five-class model that integrates text and structured information to improve its prediction accuracy and generalization ability?

RQ2: How to use SHAP to reveal the decision logic and key influencing features of the Yelp sentiment classification model?

RQ3: How to combine SHAP and confusion matrix to analyze the misclassification patterns and mechanisms of the Yelp model on fine-grained ratings?

1.5 Research Aim and Objectives

This study aims to develop and evaluate an enhanced explainable Yelp five-class sentiment prediction framework to significantly improve model performance in fine-grained sentiment classification tasks, thereby supporting enhanced accuracy and trustworthiness in user review analysis. The research objectives are clearly delineated into the following three core directions:

Obj1: To construct a five-class sentiment prediction model integrating Yelp review text semantics with business structured metadata, and evaluate its performance improvement in terms of classification accuracy and generalization ability.

Obj2: To apply the SHAP method to reveal the internal decision logic of the constructed model in the sentiment classification task, and identify key text features and business attributes that significantly influence prediction results.

Obj3: To combine SHAP interpretation results with confusion matrix analysis to deeply explore and visualize the model's misclassification patterns and their underlying mechanisms when distinguishing fine-grained ratings (especially 4-star and 5-star).

1.6 Scope of the Study

To ensure the depth and feasibility of the study, the scope of this research is clearly defined as follows:

1. Data Source and Type: Publicly available English text reviews, 1-5 star rating labels, and business metadata from the Yelp platform will be used. The study focuses on single-language (English) text and does not include multilingual processing or multimodal data such as images or audio.

2. Model Construction and Complexity: The focus is on building a five-class sentiment prediction model that integrates textual semantic features extracted by BERT with structured metadata. The model will be based on established machine learning classification algorithms and will not extend to developing entirely new end-to-end deep learning architectures (except for BERT itself) or more complex models like graph neural networks.

3. Explainability Method and Analysis Depth: The SHAP method will be primarily used to analyze the model's decision process and the influence of key features. Confusion matrix analysis will be combined to analyze and visualize misclassifications of fine-grained ratings (particularly 4-star and 5-star). The study does not involve advanced explainability techniques such as model retraining based on explanation feedback, causal inference, or generating natural language explanations.

4. Implementation Environment and Nature of Results: Experiments will be conducted using Python and standard machine learning/NLP libraries in a standard research environment such as local setup or Colab. The research outcomes are intended for theoretical validation and understanding model behavior, and do not involve large-scale system deployment, production environment applications, or user interface development.

1.7 Significance of the Research

This study endeavors to enhance the accuracy, transparency, and practical value of sentiment classification models. The expected contributions of this research are primarily manifested at the following three levels:

Firstly, on the theoretical level, this study addresses the research gaps in the current sentiment analysis field concerning "semantic-label bias," "lack of interpretability," and "underutilization of structural features." By integrating BERT embeddings with structured metadata in a five-class task and introducing SHAP for prediction explanation, the research is expected to broaden the input dimensions and output interpretation capabilities of existing semantic modeling methods, providing a theoretical demonstration and experimental basis for multi-source sentiment modeling and explainability fusion.

Secondly, on the methodological level, this study will construct a sentiment prediction model that balances performance and interpretability. Alongside evaluating classification performance, it will analyze the model's feature attention mechanisms and prediction bias distribution. This "structure-semantics-explanation" trilogy design path helps to compensate for the limitations in current research that either prioritize accuracy over interpretability or fail to integrate structural information. It also provides a reusable framework and implementation paradigm for subsequent scalable model designs, such as attention fusion, graph modeling, and domain adaptation.

Finally, on the application level, this research emphasizes fine-grained analysis of the misclassification mechanisms for easily confused sentiments in Yelp data, such as "4-star vs. 5-star" and "neutral vs. slightly negative." This analysis provides data support for business service feedback analysis, automated review monitoring, and optimization of platform rating mechanisms. The research outcomes can not only serve the internal "explanation-correction-feedback" loop construction within sentiment analysis systems but also provide explainability guarantees and practical references for enhancing the trustworthiness of platform algorithms.

In summary, this research simultaneously considers semantic depth, structural information, and explanation mechanisms in the text sentiment analysis task. It possesses not only theoretical research extension value but also practical potential in real business contexts, and is expected to provide a valuable practical sample and methodological reference for sentiment understanding research targeting real-world tasks within the data science field.

1.8 Structure of the Thesis

This research thesis is divided into five chapters, with the content arrangement for each chapter as follows:

1. Introduction is in the Chapter 1. This chapter primarily outlines the background and motivation of this study, clarifies the research questions, objectives, scope, and theoretical and practical significance, aiming to guide the reader to establish an overall understanding of the research topic.

2. Literature Review is in the Chapter 2. This chapter will review key literature in the field of sentiment analysis, with a focus on research progress in semantic modeling, structured information fusion, and model interpretability, and identify the shortcomings and gaps in existing research.

3. Research Methodology is in the Chapter 3. This chapter will detail the five-class sentiment prediction framework proposed in this study, including data preprocessing, feature fusion based on BERT and structured information, model construction and training, and implementation details of SHAP explainability analysis.

4. Experimental Design and Results Analysis is in the Chapter 4. This chapter will present the classification performance of the proposed model and, through confusion matrix and SHAP analysis, deeply explore the model's misclassification patterns and prediction logic to verify the achievement of research objectives.

5. Conclusion and Future Work is in the Chapter 5. This chapter will summarize the main conclusions and contributions of this study, discuss the limitations of the research, and provide an outlook on future research directions such as method optimization, data expansion, and explainability deepening.

1.9 Summary

This chapter systematically introduces the development background of sentiment analysis in UGC scenarios and focuses on the key challenges faced by fine-grained sentiment classification tasks on the Yelp platform, including issues such as semantic-label bias, lack of model interpretability, and insufficient utilization of structured information. Through problem analysis, this chapter clearly proposes specific research questions and research objectives, providing clear direction and motivational support for subsequent modeling and analysis work.

This chapter also clarifies the scope and limitations of this study, defining the data type, modeling complexity, boundaries for the use of explainability techniques, and the selection of the technical platform, thus ensuring that the research is sufficiently focused and executable.

Finally, this chapter provides an overview of the overall structural arrangement of this report, offering a framework guide for readers to understand the logical

development of subsequent chapters. The next chapter will conduct a systematic literature review, focusing on reviewing representative research achievements in related fields such as sentiment analysis, structured data fusion, and model interpretability in recent years, further demonstrating the research value and theoretical gaps of this study.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

Since the advent of the internet and social networks, Sentiment Analysis has evolved into the most important area of research in Natural Language Processing. Sentiment Analysis not only identifies user emotions and the direction of public opinion but also plays an important role in application areas like product recommendation, market promotion, and financial prediction. In the last several years, the availability of large-scale User-Generated Content (UGC) from diverse sources such as Yelp, Amazon, IMDB, and Twitter has greatly enriched the data resource for sentiment analysis. The expansion has also brought about new challenges in terms of requirements for finer-grained labeling, increased data variety, as well as more sophisticated business needs. Thus, it is necessary to revisit comprehensively mainstream datasets, modeling methods, explainability techniques, and misclassification analysis theories to shed light on the development trajectory of the field, comprehend state-of-the-art challenges, and outline future research directions.

This chapter will systematically elaborate on the key issues like multi-source sentiment analysis datasets, traditional and deep learning approaches, explainability analysis, and multi-class misclassification measure. Through summarizing and contrasting the existing research attainments, we aim to explore the common issues and technical challenges in the task of multi-class sentiment modeling. The chapter will ultimately formulate key unsolved research gaps with a vision of laying a theoretical foundation and practical groundwork for offering follow-up research contents and developing novel approaches.

2.2 Application Background and Motivation

In recent years, the rapid advancement of the internet and Web 2.0 technologies has driven a full-scale explosion of User-Generated Content (UGC). Users actively publish reviews about products, services, and experiences on social platforms, review websites, and forums. This UGC has become an indispensable resource for information dissemination, consumer decision-making, and business management in modern society (Almansour et al., 2022). Statistics show that over 75% of consumers actively check online reviews before making purchasing decisions, and most users consider these reviews more valuable than advice from friends and family (Almansour et al., 2022).

Among the many UGC platforms, Yelp stands out for its focus on local life services, dining, and entertainment, accumulating a vast amount of high-quality review data. What makes Yelp unique is that, in addition to user-written text reviews, it systematically integrates multi-dimensional structured information such as user ratings (star ratings), timestamps, geographical locations, business categories, price ranges, and user activity. This provides an ideal scenario and rich samples for multi-source data fusion and large-scale sentiment analysis research (Alamoudi & Alghamdi, 2021; Lak & Turetken, 2014).

However, the explosive growth of UGC reviews has also brought unprecedented challenges. Firstly, information overload is increasingly prominent, making it difficult for users and businesses to efficiently filter truly valuable review content (Almansour et al., 2022). Secondly, UGC content is highly subjective, with significant variations in expression, length, and focus; some reviews are detailed and useful, while others are off-topic or lack practical reference value (Alamoudi & Alghamdi, 2021). More complexly, the actual emotional polarity in user ratings and their text reviews is not always consistent. This "Text-Rating Review Discrepancy" (TRRD) problem has been confirmed by multiple studies, and simply using ratings as sentiment labels can mislead subsequent model training and prediction (Almansour et al., 2022; Lak & Turetken, 2014).

Facing these challenges, there's a growing urgent demand from both society and academia for "fine-grained, multi-faceted sentiment analysis." Traditional sentiment analysis methods primarily focus on overall polarity judgment. However, in real life and business decisions, users are often more concerned about specific evaluations of reviews across different aspects (e.g., taste, ambiance, service, and value for money in the catering industry). For instance, a single review might contain multi-dimensional opinions such as "the dishes were delicious but the service was poor," where a single polarity label cannot accurately reflect its complex emotional structure (Fan & Zhang, 2024; Alamoudi & Alghamdi, 2021).

Therefore, sentiment analysis of Yelp review data not only helps consumers make more informed choices but also provides crucial data support for businesses in precision marketing, reputation management, and product improvement. Simultaneously, advancing cutting-edge topics such as multi-source information fusion, fine-grained sentiment recognition, and text-rating consistency modeling has become a mainstream research direction and a hot topic in industry practice within the field of sentiment analysis (Fan & Zhang, 2024; Almansour et al., 2022).

2.3 Overview and Comparison of Main Sentiment Analysis Datasets

For the diverse tasks of sentiment analysis, academia and industry have accumulated a variety of high-quality public datasets, supporting the continuous evolution of models and methods.

2.3.1 Yelp Dataset

The Yelp dataset is one of the most widely used real-world review big data in the current research fields of sentiment analysis and recommendation systems. This dataset was first launched by Yelp and has been continuously updated and upgraded. It currently contains millions of user reviews, star ratings, user information, and business attributes from industries such as catering, leisure, and entertainment

worldwide. Yelp data is not only highly structured but also includes rich metadata such as free text, timestamps, geographical information, price ranges, and tags, making it very suitable for multi-source data modeling, fine-grained sentiment analysis, and explainability research (Alamoudi & Alghamdi, 2021). Yelp's labeling system mainly uses 1–5 star ratings, supporting multi-class classification and regression analysis. It has high representativeness and application value in cutting-edge research such as multi-aspect polarity modeling, star-text inconsistency, and UGC multimodal mining (Fan & Zhang, 2024).

The Yelp Academic Dataset can be downloaded for free after registration on its official website (<https://www.yelp.com/dataset>) for academic or non-commercial research use only. The official website will update the data version from time to time, which is convenient for researchers to reproduce and compare methods (Alamoudi & Alghamdi, 2021).

2.3.2 Amazon Review Dataset

The Amazon Review dataset is another large-scale, broadly diversified, general-purpose sentiment analysis dataset. This dataset aggregates tens of millions of user reviews and rating information from various industries on the Amazon platform (such as books, digital products, apparel, home goods, etc.). It features detailed structured data including product IDs, categories, text, ratings (1–5 stars), timestamps, and user behavior. Compared to Yelp, the Amazon Review dataset exhibits stronger cross-category and spatio-temporal characteristics, supporting more complex applications like domain transfer, product recommendations, and user behavior modeling (Katić & Milićević, 2018). Its primary challenges include uneven review distribution, a higher prevalence of fake reviews, and numerous outliers. Nevertheless, its sheer scale and diversity offer significant value for academic and engineering research (He & McAuley, 2016).

The Amazon review dataset can be obtained through various channels, such as Kaggle, AWS Public Datasets, or researchers' homepages. The most classic "Amazon

Product Data" is continuously maintained by the McAuley team (<https://nijianmo.github.io/amazon/index.html>), offering convenient downloads, clear categorization, and applicability to various specific tasks (He & McAuley, 2016).

2.3.3 IMDB Movie Review Dataset

The IMDB Movie Review dataset is one of the earliest and most standard benchmarks in the field of text sentiment analysis. First compiled and released by Maas et al., this dataset primarily consists of free text movie reviews with positive and negative labels, making it suitable for binary sentiment classification and text feature extraction experiments. The IMDB dataset has a simple structure, typically features longer text lengths, and generally does not include additional user or product metadata. Its main advantages are clean data and accurate labeling, which makes it ideal for baseline testing of new algorithms and preliminary sentiment analysis research (Maas et al., 2011). However, compared to Yelp and Amazon datasets, its extensibility is limited for tasks such as multi-class classification, fine-grained analysis, and multi-source fusion.

The IMDB sentiment analysis dataset can be downloaded for free from the Stanford NLP team's official website (<https://ai.stanford.edu/~amaas/data/sentiment>) required. Its standardized format and clear labels make it highly suitable for academic experiments (Maas et al., 2011).

2.3.4 X (Twitter) Dataset

The X (Twitter) dataset is an important benchmark for social media text analysis and sentiment computing. This type of dataset primarily consists of tweets, which are short, information-dense texts covering various topics such as politics, entertainment, current events, and product feedback. X sentiment analysis primarily focuses on tasks like positive-negative-neutral three-class classification, topic extraction, and network event evolution, with common datasets including

Sentiment140 and SemEval. X data is characterized by strong timeliness, large volume, diverse language styles, and rich emojis. However, it also faces challenges such as high text noise, complex implicit emotional expressions, and frequent use of irony and puns (Wang et al., 2022; Qi & Shabrina, 2023). X data is highly suitable for cutting-edge research in multi-task learning, sentiment trend mining, and online emotion prediction.

Typical X (Twitter) sentiment datasets like Sentiment140 and SemEval can be downloaded for free from project homepages or the ACL data platform. However, the original tweet content needs to be scraped via the X API (requiring a developer account). Due to platform policies and tweet ID validity, data collection has certain technical barriers, but it is highly open and globally applicable (Wang et al., 2022; Qi & Shabrina, 2023).

2.3.5 Comparative Analysis of Four Major Datasets

In order to more intuitively demonstrate the similarities and differences in structure, label types, and applicable scenarios among the four major sentiment analysis datasets, their core features are summarized in the table below.

Table 2.3.5 Overview of Main Public Sentiment Analysis Datasets

Date set	Data Size	Domain	Label Type	Meta data	Typical Text Length	Representative Applications	Main Challenges	References
Yelp	Millions	Local Life/Dining	1–5 stars	Rich	Medium	Multiclass, fine-grained sentiment,	Text-rating inconsistency, high	Alamoudi & Alghamdi

						explainability	subjectivity	(2021); Fan & Zhang (2024)
Amazon	Tens of millions	All product categories	1–5 stars	Rich	Medium–Long	Recommendation, domain adaptation, multi-source analysis	Uneven distribution, fake/extreme reviews	Katić & Milićević (2018); He & McAuley (2016)
IMDB	Tens to hundreds of thousands	Movies	Binary	Simple	Long	Binary classification, feature extraction, baseline testing	Lack of diversity, limited scalability	Maas et al. (2011)
X (Twitter)	Millions	Social domain	Three/multiclass	General	Short	Opinion mining, event tracking, NLP benchmarks	High noise, sarcasm, complex expressions	Wang et al. (2022); Qi & Shabrina (2023)

As summarized above, mainstream public datasets like Yelp, Amazon, IMDB, and X (Twitter) each have their own strengths in terms of data scale, domain focus, labeling systems, and metadata richness, providing a solid foundation for sentiment analysis research. Selecting a dataset with matching characteristics for a specific research question is crucial for model performance, generalization ability, and the effectiveness of the final application (Katić & Milićević, 2018; He & McAuley, 2016; Wang et al., 2022).

Beyond these standard datasets, dynamically collecting raw reviews or news data for specific scenarios through techniques like web scraping has also become an important trend. This approach helps improve the timeliness and task specificity of the data, and it can enhance researchers' comprehensive abilities in data engineering and practical business scenarios (Kaur, 2022).

Therefore, future research will comprehensively consider the applicability of existing public datasets and the feasibility of self-collected data, flexibly designing data acquisition and preprocessing workflows.

2.4 Overview and Comparative Analysis of Sentiment Analysis Methods

As sentiment analysis tasks continue to evolve, researchers have proposed various modeling approaches. Based on different technical routes and theoretical foundations, mainstream methods can be broadly categorized into three types: traditional machine learning methods, classic methods specifically designed for sentiment analysis, and deep learning and pre-trained model methods. A systematic review of each will be provided below.

2.4.1 Traditional Machine Learning Approaches

1. Naive Bayes (NB)

Naive Bayes (NB) is a probabilistic classifier widely used for text sentiment classification due to its simplicity, efficiency, and effectiveness with high-dimensional, sparse features. The core assumption of NB is that features (such as words or n-grams in a document) are conditionally independent given the class. In sentiment analysis, NB calculates the posterior probability of each sentiment class given the observed features, assigning the class with the highest probability as the final label (Arya et al., 2022; Fransisca et al., 2021; Ghatora et al., 2024; Ramasamy et al., 2023; Das et al., 2023).

The most representative classification formula for Naive Bayes in text sentiment analysis is as follows:

$$\hat{y} = \arg \max_{y \in Y} P(y) \prod_{i=1}^n P(x_i | y)$$

where \hat{y} is the predicted sentiment label, Y is the set of possible sentiment classes (e.g., positive, negative, neutral), x_i represents the i th feature (word) in the text, $P(y)$ is the prior probability of class y , and $P(x_i | y)$ is the conditional probability of observing feature x_i given class y (Ghatora et al., 2024; Fransisca et al., 2021).

Naive Bayes has also been found to work well with short-text and moderately sized datasets for languages and platforms (Arya et al., 2022; Das et al., 2023). Its principal limitation, though, is the strong independence assumption, which can be untrue for natural language; thus, its performance deteriorates in the presence of highly correlated features or subtle/complicated sentiment expressions (Fransisca et al., 2021; Das et al., 2023; Ghatora et al., 2024).

2. Support Vector Machine (SVM)

Support Vector Machine (SVM) is a discriminative classifier that seeks to determine the best separating hyperplane with maximum margin between classes in a

high-dimensional feature space. SVM is very efficient for text classification problems like sentiment analysis, particularly in the case of sparse and high-dimensional data (Han et al., 2020; Das et al., 2023; Ghatora et al., 2024; Singh et al., 2022; Benarafa et al., 2024).

The SVM classification decision function used in sentiment analysis is given by:

$$f(z) = \text{sgn} \left(\sum_{i=1}^l \alpha_i y_i K(x_i, z) + b \right)$$

where z is the feature vector of the text to be classified, x_i are support vectors from the training set, y_i are their class labels, α_i are the learned weights, $K(x_i, z)$ is the kernel function measuring similarity, b is the bias term, and sgn denotes the sign function which assigns the final class (Benarafa et al., 2024; Han et al., 2020).

SVM has demonstrated superior performance compared to Naive Bayes and other traditional models, particularly for binary and multi-class sentiment classification tasks with short texts, such as product reviews and tweets (Das et al., 2023; Ramasamy et al., 2023; Ghatora et al., 2024). Advanced versions using kernel tricks (such as RBF, polynomial, or semantic kernels) further enhance its ability to capture non-linear relationships and implicit sentiment aspects (Benarafa et al., 2024; Han et al., 2020). However, SVM can be computationally intensive for very large datasets and requires careful tuning of hyperparameters and kernel selection (Ghatora et al., 2024; Ramasamy et al., 2023).

3. K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) is a non-parametric, instance-based learning algorithm commonly used as a baseline for sentiment analysis tasks. Its core principle is to classify an unseen text sample by the majority sentiment label among its k closest

neighbors in the feature space, based on a similarity or distance measure (Abo et al., 2021; Saady et al., 2022).

KNN requires no explicit training phase; instead, it stores all training samples and classifies new samples on demand. Typical workflows in sentiment analysis include text vectorization (e.g., TF-IDF), normalization, and the application of KNN for multi-class sentiment prediction. KNN is valued for its simplicity, interpretability, and ease of implementation, especially for small to medium-sized datasets where category boundaries are distinct (Saady et al., 2022).

However, KNN suffers from several limitations: its computational cost increases rapidly with dataset size, it is sensitive to the curse of dimensionality and noise, and its performance strongly depends on the choice of distance metric and value of k (Abo et al., 2021). In empirical studies on Arabic sentiment analysis and mobile app review classification, KNN often achieved lower accuracy and F1-scores compared to advanced models such as Random Forest, Logistic Regression, and deep learning approaches, but it remains a valuable baseline for algorithm comparison and hybrid model integration (Saady et al., 2022; Abo et al., 2021).

4. Logistic Regression (LR)

Logistic Regression (LR) is a popular linear classifier used for sentiment analysis, particularly effective for large-scale and high-dimensional text data. For multiclass sentiment classification tasks (e.g., Yelp 1–5 stars), LR is typically extended to multinomial logistic regression using the softmax function, modeling the probability that a text belongs to each of the K sentiment classes:

$$P(\mathbf{y} = \mathbf{k} | \mathbf{X}) = \frac{\exp(\beta_{k,0} + \sum_{i=1}^n \beta_{k,i} \mathbf{x}_i)}{\sum_{j=1}^K \exp(\beta_{j,0} + \sum_{i=1}^n \beta_{j,i} \mathbf{x}_i)}$$

where $P(\mathbf{y} = \mathbf{k}|\mathbf{X})$ is the probability of assigning class \mathbf{k} to text sample \mathbf{X} , $\beta_{k,0}$ and $\beta_{k,i}$ are the bias and feature weights for class k , and \mathbf{x}_i are the text features (Wenping Wang et al., 2023; Singh & Jaiswal, 2023).

Empirical findings indicate that multinomial logistic regression, particularly when coupled with vectorization methods like TF-IDF, is competitive in multiclass sentiment analysis tasks like Yelp and Twitter reviews. It is as good as, if not superior to, more sophisticated models with the added advantage of fast inference speed and interpretability (Wenping Wang et al., 2023; Padhy et al., 2024). Yet, its linear modeling capacity could restrict its effectiveness in capturing sophisticated patterns of sentiment when feature interactions and context are important (Singh & Jaiswal, 2023).

5. Decision Trees (DT)

Decision Trees (DTs) are popular non-parametric supervised multiclass sentiment analysis models appreciated for their interpretability and ability to deal with categorical and numerical features. For sentiment classification, DTs learn a hierarchical tree model by recursively dividing the feature space—e.g., bag-of-words, TF-IDF, or word embedding vectors—according to feature values maximising a split criterion.

At each internal node, the algorithm chooses a feature and a threshold that give the best discrimination among sentiment classes based on measures like information gain or the Gini index (Dandash & Asadpour, 2023; Jain et al., 2023).

An example of a split criterion representative for the CART decision tree algorithm is the Gini index:

$$Gini(D) = 1 - \sum_{k=1}^K p_k^2$$

where $Gini(D)$ is the impurity of the node containing dataset D , K is the number of sentiment classes, and p_k is the proportion of samples belonging to class k at that node. The algorithm selects the split that results in child nodes with minimal weighted Gini impurity.

Decision Trees are attractive for sentiment analysis due to their straightforward structure and ability to visualize the decision-making process—helping reveal which words or features most strongly affect the sentiment classification. Empirical results demonstrate that DTs, when combined with appropriate feature representations, can perform competitively on multiclass sentiment analysis tasks. For example, Dandash & Asadpour (2023) report that DTs, applied to Arabic social media sentiment classification with bag-of-words and TF-IDF features, achieved accuracies ranging from 22% to 38% across various multiclass settings. Jain et al. (2023) highlight DTs as a key component in multimedia sentiment analysis pipelines, and in broader benchmark comparisons, decision trees often serve as interpretable baselines or as base learners in ensemble methods (such as Random Forests and Gradient Boosted Trees) (Dandash & Asadpour, 2023; Jain et al., 2023).

However, DTs are prone to overfitting in high-dimensional and sparse feature spaces, especially common in text data, and their performance may be surpassed by more robust algorithms like SVM, logistic regression, or deep learning models. Nevertheless, their transparency and ability to natively handle multiclass tasks make them valuable tools for both standalone and ensemble sentiment classification systems.

6. Random Forest (RF)

Random Forest (RF) is an ensemble approach whereby a huge number of decision trees are trained on the train data at train time and their predictions aggregated together for improving classification accuracy and generalization. RF grows each tree on a bootstrapped version of the train data and uses a random subset of features at each

split, making the ensemble stronger and less vulnerable to overfit compared to single decision trees (Dandash & Asadpour, 2023; Jain et al., 2023). The final sentiment class for a given instance is determined based on the voting decision of all of the forest's trees:

$$\hat{y} = \text{majority_vote} \{h_t(\mathbf{x}) \mid t = 1, \dots, T\}$$

where \hat{y} is the sentiment prediction for input vector \mathbf{x} , $h_t(\mathbf{x})$ is the t -th tree prediction and T is the total number of trees.

Random Forest has been found suited well for sentiment classification on text-based high-dimensional bag-of-words, TF-IDF, or n-grams. Empirical work demonstrates RF working reliably on multi-class sentiment analysis. Dandash & Asadpour (2023), for instance, established the performance of RF superior compared to single decision trees, KNN, and a number of linear classifiers on Arabic sentiment corpora from Twitter and Facebook containing as high as 40.54% accuracy from character-level TF-IDF features. Jain et al. (2023) also cite RF as a competitive and popular baseline for sentiment analysis and multimedia pipelines and report competitive performance on a variety of benchmark datasets.

Despite these strengths, Random Forest can be computationally demanding as the number of trees and features increases, and while it supports feature importance analysis, its decision process is less transparent than that of a single tree. Nonetheless, RF remains one of the most popular and effective methods for multiclass sentiment analysis, valued for its predictive power and reliability in practical applications (Dandash & Asadpour, 2023; Jain et al., 2023).

7. Model Comparison

(a) Traditional machine learning models commonly used in sentiment analysis each have their own characteristics. The table below compares the main

advantages, limitations, typical application scenarios, and relevant references of these models.

Table 2.4.1.7 Comparison of Traditional Machine Learning Models for Sentiment Analysis

Model	Advantages	Limitations	Typical Use Case	References
Naive Bayes	Fast, interpretable, handles sparse data well	Independence assumption, lower accuracy on complex data	Short/simple texts, baseline	Arya et al., 2022; Ghatora et al., 2024
Support Vector Machine	High accuracy, strong generalization	Computationally intensive, needs tuning	Binary and multiclass text classification	Han et al., 2020; Benarafa et al., 2024
K-Nearest Neighbors	Simple, no explicit training	Not scalable, sensitive to noise	Small datasets	Abo et al., 2021; Saudy et al., 2022
Logistic Regression	Efficient, interpretable, good with multiclass	Linear, limited with complex non-linear patterns	Large, multiclass datasets	Wenping Wang et al., 2023; Singh & Jaiswal, 2023
Decision Trees	Interpretable, supports multiclass	Overfitting, less robust on large/sparse data	Visualization, analysis	Dandash & Asadpour, 2023; Jain et al., 2023
Random Forest	High accuracy, robust, reduces overfitting	Less interpretable, resource-intensive with many trees	Complex, multiclass tasks, feature importance	Dandash & Asadpour, 2023; Jain et al., 2023

As evident from the table, different models vary in terms of accuracy, training efficiency, and applicable scenarios. Naive Bayes and Logistic Regression are suitable

for high-dimensional sparse text due to their computational efficiency, but they have limitations in modeling complex relationships. SVM and Random Forest excel in classification performance but demand higher computational resources. In practical applications, it's crucial to select models judiciously, considering data characteristics and project requirements to achieve a balance between performance and efficiency.

2.4.2 Sentiment-specific Classic Methods

Common dictionary and rule-based methods used in sentiment analysis include SentiWordNet, VADER, and TextBlob. These methods don't rely on large amounts of manually labeled data; instead, they determine the overall sentiment polarity of a text through pre-built sentiment lexicons combined with simple score aggregation or rule-based judgments (Taboada et al., 2011; Qi & Shabrina, 2023).

1. Principles of Mainstream Dictionary-Based Methods

- SentiWordNet assigns positive, negative, and neutral scores to each word or phrase based on WordNet's synsets. It is suitable for general English text but has limitations in handling internet slang and emoticons (Nursal et al., 2025).
- VADER is specifically designed for social media, with its lexicon covering internet slang, emojis, and colloquialisms. It also incorporates rules (e.g., negation, exclamation, lexical emphasis) to adjust sentiment intensity, making it particularly effective for short texts like those found on Twitter and forums (Qi & Shabrina, 2023; Nursal et al., 2025).
- TextBlob primarily scores each word or phrase based on its lexicon, then calculates the average overall polarity. It's suitable for general texts and product reviews (Qi & Shabrina, 2023).

2. Empirical Comparisons and Application Characteristics

A recent study (Nursal et al., 2025) compared the performance of WordNet, SentiWordNet, TextBlob, and VADER on Malaysian high-rise residential forum and Google review data, revealing the following:

- All lexicon-based methods tend to identify more positive sentiment, but VADER performs better in identifying negative and neutral sentiment, achieving the highest overall classification accuracy (78%) and a recall rate of up to 90%.
- SentiWordNet and WordNet have broad coverage but limited adaptability to slang and new words. They are susceptible to social media noise, resulting in slightly lower accuracy than VADER and TextBlob.
- TextBlob's overall performance is moderate, suitable for general English text analysis, but it's somewhat limited when handling informal text and nuanced emotions.

Moreover, common advantages of lexicon-based methods include simple implementation, computational efficiency, ease of interpretation, independence from training samples, suitability for cold-start and low-resource scenarios, or as a baseline for machine learning models (Ghatora et al., 2024; Nursal et al., 2025). However, they also have clear shortcomings: difficulty in recognizing complex expressions such as sarcasm, negation, ambiguity, and spelling errors, and limitations in lexicon coverage can affect adaptability to new domains (Nursal et al., 2025).

2.4.3 Deep Learning and Pre-trained Language Models

2.4.3.1 Basic Deep Learning Methods

In recent years, deep learning techniques have made significant strides in the field of sentiment analysis. Traditional machine learning methods rely on manual feature extraction, whereas deep learning models can automatically learn complex

semantic and temporal features from raw text, greatly improving the accuracy of sentiment classification (Wu et al., 2022; Jin, 2023).

1. Convolutional Neural Networks (CNN)

RNNs and their extensions (LSTM, GRU, and BiLSTM) can capture contextual dependency in a sequence of text. The vanishing and exploding gradient problem of traditional RNNs when dealing with long sequences can be overcome by LSTM and GRU through gate mechanism and thus they are better suited for capturing long texts and semantically highly consistent contexts. BiLSTM also employs bidirectional information and thus inherits their ability to capture contextual sentiment (Golubeva & Loukachevitch, 2021; Wu et al., 2022).

2. Recurrent Neural Networks (RNN) and Variants

The researchers have proposed a series of ensemble architecture models for better sentiment analysis performance. As an example, CNN-LSTM combines the advantage of CNN in local feature extraction and the advantage of LSTM in temporal dependency modeling so as to achieve a good sense of sentiment of text (Wu et al., 2022). Similarly, multi-level models like Co-LSTM and Two-Level LSTM enhance the advantage of the model in grasping complex emotional expressions (e.g., inversions and negations) through the addition of sentiment lexicons or polarity reversal (Wu et al., 2022; Jin, 2023). Meanwhile, the integration of LSTM and GRU has also seen fruitful applications in a plethora of areas ranging from cryptocurrency and finance to healthcare (Jin, 2023).

3. Ensemble and Hybrid Models

The researchers have proposed a series of ensemble architecture models for better sentiment analysis performance. As an example, CNN-LSTM combines the advantage of CNN in local feature extraction and the advantage of LSTM in temporal dependency modeling so as to achieve a good sense of sentiment of text (Wu et al., 2022). Similarly, multi-level models like Co-LSTM and Two-Level LSTM enhance

the advantage of the model in grasping complex emotional expressions (e.g., inversions and negations) through the addition of sentiment lexicons or polarity reversal (Wu et al., 2022; Jin, 2023). Meanwhile, the integration of LSTM and GRU has also seen fruitful applications in a plethora of areas ranging from cryptocurrency and finance to healthcare (Jin, 2023).

2.4.3.2 Pre-trained Language Models

Over the last few years, sentiment analysis has also seen a big boost through pre-trained Transformer language models. Pre-trained models like BERT can capture more semantic and contextual relationships through unsupervised pretraining on a large corpus of texts and hence significantly enhance the performance of downstream applications like sentiment classification (Devlin et al., 2019).

1. BERT and its Variants

BERT employs a multi-layer bidirectional Transformer encoder capable of simultaneously capturing both left and right context words of a given text. BERT introduced two-stage unsupervised pre-training tasks: Masked Language Model (MLM) and Next Sentence Prediction (NSP). BERT significantly enhanced the precision and generalization capacity of sentiment classification via its "pre-training + fine-tuning" mechanism (Devlin et al., 2019).

BERT and variants thereof (e.g., RoBERTa, DistilBERT, and BERTweet) also demonstrated high cross-linguistic and cross-domain generalizability. Multilingually trained models like RuBERT and GreekBERT also demonstrated excellent performance in sentiment analysis for Russian and Greek languages (Syrigka et al., 2023; Golubeva & Loukachevitch, 2021).

2. Multi-Model Fusion and Architectural Innovations

With increasing application requirements, researchers begin to combine BERT-type models with traditional deep learning models (e.g., CNN, BiLSTM,

ResNeXt, etc.) for enhanced feature extraction capacity and better performance of the model. As a case in point, triple fusion approach RoBERTa-ResNeXt-BiLSTM results in higher customer review sentiment analysis precision compared to standalone models (Farah et al., 2024). Dual-channel deep classifier (DJC) of RoBERTa and BERT also can effectively handle imbalanced data and multi-category sentiment and improve the generalization capacity and stability of the model (Zhang et al., 2024).

3. Comparative Analysis and Applications of Various Pre-trained Models

Recent research indicates BERT and BERT variant models (e.g., RoBERTa, GPT-2, XLNet) performing better than traditional machine learning and shallow deep-learning models on various publicly available datasets such as IMDB, Yelp, Twitter, and clinical medical discussions (Syrigka et al., 2023). Amongst them, RoBERTa excels at long-text and fine-grained polarity responses, BERT would be naturally fit for multi-lingual contexts, and the generative models GPT-2 and XLNet perform well on emotional reasoning and text generation tasks.

2.5 Overview of Explainability Methods in Sentiment Analysis

2.5.1 The Evolution of Explainability Techniques

With deep learning models increasingly used in sentiment analysis, the "black-box" phenomenon has become a prominent concern recently. In other words, the models can make predictions but cannot as clearly explain the decision-making foundations behind their predictions, which hinders their application in high-risk sectors and actual business operations. Consequently, a surge in explainability (Explainability/Interpretability) research has emerged in the AI and NLP fields, driving the birth of various interpretability techniques (Ghasemi & Momtazi, 2023). On one hand, as a task closely related to subjective user experience, sentiment analysis results often influence business decisions, policy evaluations, and even financial trends, making it particularly crucial to "make users trust why the model made a certain sentiment judgment" (Rizinski et al., 2024). On the other hand, legal regulations (such

as the EU GDPR) also mandate that some AI systems must possess traceability and explainability for their results, which has further driven the development of relevant methods (Tutek & Šnajder, 2022).

2.5.2 Mainstream Explainability Methods Categorization

1. Local Interpretable Model-Agnostic Explanations (LIME)

LIME is a post-hoc, model-agnostic explanation method that can reveal the influence of features on a single prediction by fitting a simple, interpretable linear model through local perturbations of the input data. LIME is a post-hoc, model-agnostic explanation technique that is able to disclose the feature contribution to an individual prediction by learning a simple, interpretable linear model via local perturbations of the input data. LIME is widely applied in text sentiment analysis, particularly for explaining deep models or hybrid methods, with the aim of making it clear for the user "which words/segments weighed most in the model's decision" in brief-text contexts like Twitter and product reviews (Lovera et al., 2021).

2. SHapley Additive exPlanations (SHAP)

SHAP, derived from Shapley values in game theory, calculates the marginal contribution of each input feature to the output result. Over the past several years, SHAP has been utilized to interpret the discrimination process of sophisticated models such as Transformer and BERT for sentiment analysis in particularly high-stakes fields such as finance and healthcare. It has been established that SHAP can be leveraged to generate "explainable lexicons" (e.g., XLex) automatically with great gain on interpretability and some degree of performance (Rizinski et al., 2024). SHAP is one of the most widely used interpretability techniques for text classification and sentiment analysis nowadays.

3. Attention Mechanism

Attention mechanism is not just one of the most significant technologies for enhancing model performance but also a natural approach for neural network interpretability. With the visualization of attention weights, it is possible to see "on which words/phrases the model paid attention" when discriminating sentiment, giving partial explainable hints for the model decision-making process. In recent years, a large body of research has been dedicated to improving the "fidelity" and "plausibility" of attention explanations, and has proposed enhancing the consistency between attention and human cognition through regularization (Tutek & Šnajder, 2022; Ghasemi & Momtazi, 2023).

4. Explainable Visualization

Regardless of whether it's LIME, SHAP, or the Attention mechanism, their results can ultimately be presented intuitively through visualization (e.g., heatmaps, keyword highlighting, feature weight maps), helping users understand the model's internal logic and increasing trust (Lovera et al., 2021).

2.5.3 Representative Applications of Explainable Methods in Sentiment Analysis

In practical sentiment analysis tasks, different explainable methods have been widely applied to various models and scenarios:

- **Transformer and SHAP Integration:** In financial sentiment analysis, researchers leveraged a Transformer model combined with SHAP to generate a domain-specific "explainable dictionary." This approach addressed the challenges of maintaining traditional manual dictionaries and improved both model performance and transparency (Rizinski et al., 2024).
- **LIME for Explaining Deep Models:** In Twitter sentiment analysis, LIME was employed as an interpretability tool for a hybrid model (knowledge graph + deep learning). It effectively revealed the classifier's decision basis on

individual samples, enhancing the model's traceability and user trust (Lovera et al., 2021).

- **Exploring Attention Mechanism Interpretability:** It has also been established that even if attention mechanisms can make decisions of neural nets understandable, their own interpretability must be enhanced through regularization and alignment according to human annotation so as not to be disrupted by "spurious attention distribution" (Tutek & Šnajder, 2022; Ghasemi & Momtazi, 2023).
- **Explainable Multimodal Sentiment Analysis:** Multimodal and cross-lingual sentiment analysis has also been enhanced by combining a number of techniques involving the integration of hybrid attention models and counterfactual explanations in an effort to enhance explainability in models. The model can be applied in a variety of low-resource languages and hard-to-reach scenarios (Ghasemi & Momtazi, 2023).

2.6 Misclassification Analysis and Evaluation of Multi-class Sentiment Modeling

2.6.1 Multi-class Misclassification Theory and Confusion Matrix

For multi-class sentiment analysis tasks, the confusion matrix represents a critical tool for observing and pinpointing misclassification effects. Confusion in multi-class classification is more complex compared to binary classification and happens in real application contexts in which intensity between opinions overlaps or boundaries between classes lose their distinctive quality. In a rating system where five stars will be given as a restaurant rating system, models confound rating scores between 4- and 5- or 3-stars. Confounding between adjacent classes can be visualized through the confusion matrix so as to pinpoint the vulnerabilities of the model precisely (Rizinski et al., 2024). The confusion matrix not only facilitates measurement

of overall performance but also permits a nuanced analysis of each sentiment class identifiability and pitfalls common to each class.

Taking it a step further, the confusion matrix offers a stable basis for backtracking misclassifications and model refinement. For instance, examining regions of confusion between certain categories allows one to explore further the reasons for misclassification by taking into account sample text content and feature distribution. Some authors have suggested that the combination of confusion matrix with explainability analysis tools (such as LIME and SHAP) is able to better reconstruct the model's decision logic, offering a theoretical framework for the local and global optimization of sophisticated models (Lovera et al., 2021). In real business situations, the confusion matrix is also applied in conjunction with business objectives to identify "critical misclassifications" that most impact end decisions.

2.6.2 Types, Causes, and Evaluation Metrics of Misclassification

Types of misclassifications in multi-class sentiment classification consist primarily of nearby label misclassifications (e.g., 4-stars vs. 5-stars), extreme label confusions (e.g., 1-stars vs. 5-stars), and neutral vs. positive or negative polarity confusions. The causes of misclassifications tend to be based on the inherent subjectivity of data, ambiguous label meanings, highly imbalanced class distributions and complex contexts and words. Unless the model itself becomes sensitive enough toward boundary examples and minority events, it will also highly likely be misclassified as well. These issues notably manifest in real applications like fine-grained sentiment polarity classification and cross-platform multi-domain analysis (Ghatora et al., 2024).

For resolution of such misclassifications, the academic community uses a set of performance evaluation indices for the evaluation of a model's performance from a holistic viewpoint. In addition to extensively used Accuracy, Macro/Micro F1-score, Weighted F1, Recall, and Kappa coefficient, performance measurement in multiple aspects can also utilize ROC-AUC. In the specific situation of class imbalance and

complex task intensity, concurrent examination of F1-score and confusion matrix can indicate clearly the sensitivity and error points of a model for each class and provide a scientific reference for subsequent adjustment of the model (Lovera et al., 2021; Han et al., 2020).

2.6.3 Typical Countermeasures and Frontier Methods

To reduce multi-class sentiment analysis misclassification rates, researchers have proposed various solutions both at data and model levels. At the data level, undersampling/oversampling, text augmentation, and pseudo-labeling are common practiced techniques to minimize classification bias due to class imbalance. At the model level, incorporating class-weighted loss functions, hierarchical classification model architectures, and model ensembles have significantly enhanced the ability of the model to distinguish minority classes and boundary samples. These methods have been widely shown to perform well on real-world multi-class tasks such as Yelp and Twitter (Lovera et al., 2021; Ghatora et al., 2024).

In the past few years, with the pace of explainable AI (XAI) technologies' development accelerating, researchers have begun applying explanation tools like LIME and SHAP to misclassification analysis and model diagnosis. By visualizing the decision-making of the model and emphasizing high-weight features, developers can identify semantic vulnerabilities in specific misclassified instances, and as a result, data annotation, feature engineering, and model architecture can be optimized in a targeted way (Ghasemi & Momtazi, 2023). Additionally, sophisticated techniques such as adversarial training and knowledge graph feature fusion are being used more and more to enhance the robustness and generalization capacity of multi-class models, opening the way for highly reliable applications of sentiment analysis systems in complex real-world environments (Tutek & Šnajder, 2022).

2.7 Research Gaps

Despite numerous advancements in the field of sentiment analysis, there remain several research gaps in multi-class sentiment modeling and explainability analysis based on large-scale user review data (e.g., Yelp):

First, the misclassification problem in multi-class fine-grained sentiment modeling remains unsystematically addressed. Existing models in multi-category tasks, such as 1–5 star ratings, commonly exhibit phenomena like neighboring label confusion (e.g., 4-star vs. 5-star, 2-star vs. 3-star) and extreme misclassifications, which hinder practical business applications. Most mainstream research focuses on overall accuracy, with insufficient in-depth exploration of the causes of specific misclassification types, fine-grained visualization of confusion matrices, and the impact of typical confusions on subsequent decisions (Ghatora et al., 2024; Rizinski et al., 2024).

Second, the integration and application of misclassification explainability tools and methods in multi-class sentiment analysis are limited. Although explainability techniques like LIME, SHAP, and Attention have been used for single-sample or local explanations, there is currently a lack of mature engineering practices and case studies that combine these methods with multi-class misclassification analysis to form a systematic "misclassification attribution-optimization-visualization" process (Ghasemi & Momtazi, 2023; Lovera et al., 2021).

Third, there is insufficient research on explainability assisted by multi-source feature fusion and metadata. Current sentiment analysis models largely rely on primary textual features. Relevant research and empirical cases are relatively scarce concerning the synergistic explanation of structured metadata, such as user attributes, business categories, and temporal information, with text features. This makes it difficult for models to fully explain the causes of misclassifications and suggest optimization directions in real-world diverse scenarios (Rizinski et al., 2024).

Finally, for multi-class sentiment analysis on specific platforms (e.g., Yelp) facing real-world challenges such as label subjectivity and uneven data distribution, there is a lack of an explainability-driven misclassification analysis and system

improvement framework. How to combine confusion matrices, local and global explanation methods, and diverse data features to achieve traceable and actionable model optimization remains a pressing research problem (Ghatora et al., 2024).

2.8 Summary

In summary, the field of sentiment analysis has made significant progress in areas such as dataset construction, algorithm optimization, explainability, and misclassification analysis. From the diversity of mainstream public datasets to the continuous evolution of traditional and deep learning methods, and the application and integration of explainability tools like LIME, SHAP, and Attention, existing research has laid a solid foundation for improving the performance and transparency of sentiment analysis models. However, for large-scale multi-class scenarios like Yelp, there are still prominent issues such as the complexity of misclassification types, the loose integration of explainability with misclassification analysis, and the insufficient synergistic explanation of multi-source features.

This chapter clarifies the main pain points and development bottlenecks in current sentiment analysis research through a literature review. It systematically sorts out typical methods and their advantages and disadvantages, and further summarizes the key directions to be broken through in the Research Gaps section. Subsequent chapters will focus on the research proposals and methodological innovations proposed to address the aforementioned gaps, striving to achieve high performance and strong explainability in multi-class sentiment modeling, and to provide strong support for practical business needs and academic development.

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:

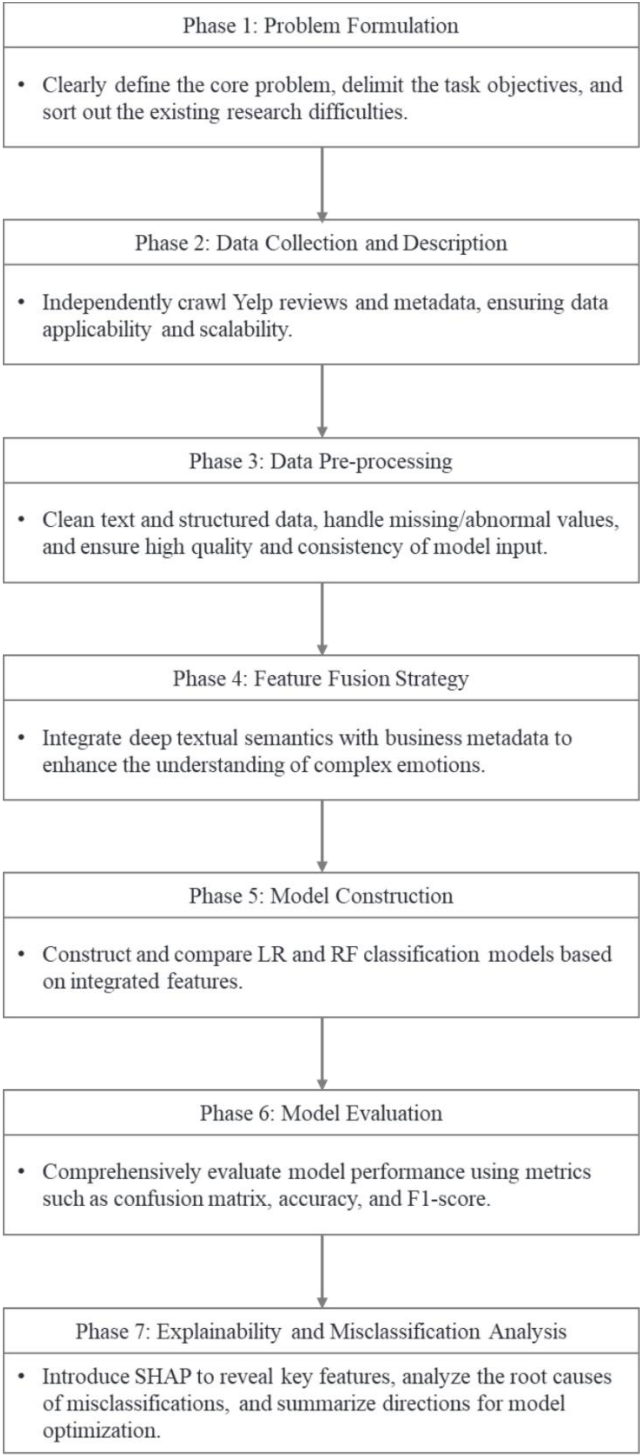


Figure 3.1 Framework diagram of research methodology workflow

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.

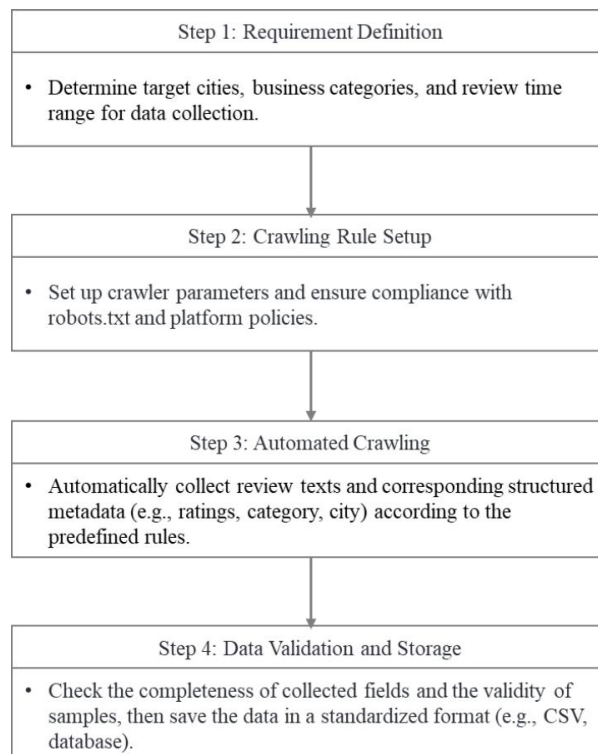


Figure 3.3 Data collection flowchart

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)**

- (b) 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)**

- (c) 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.

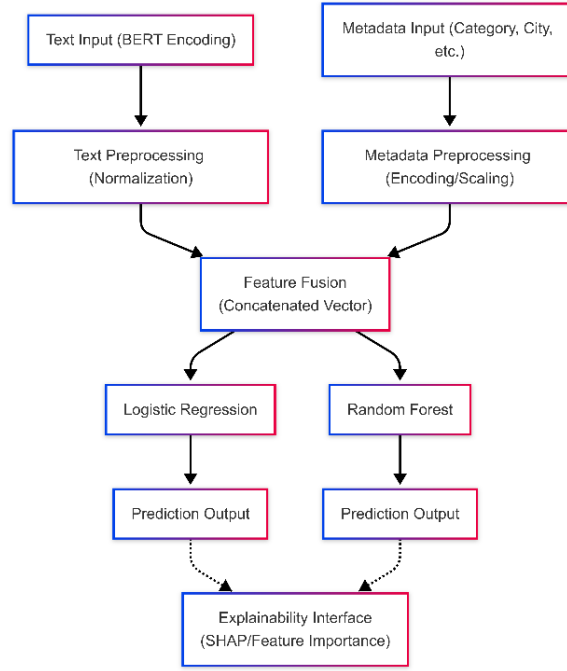


Figure 3.4 The Vertical Architecture of Multi-source Sentiment Classification Model

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)**

- (d) 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**

- (e) 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**

- (f) 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**
 - (g) 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**
 - (h) 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**
 - (i) 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**
 - (j) 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**
 - (k) 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**
 - (l) 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**

- (m) 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**

- (n) 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.

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.

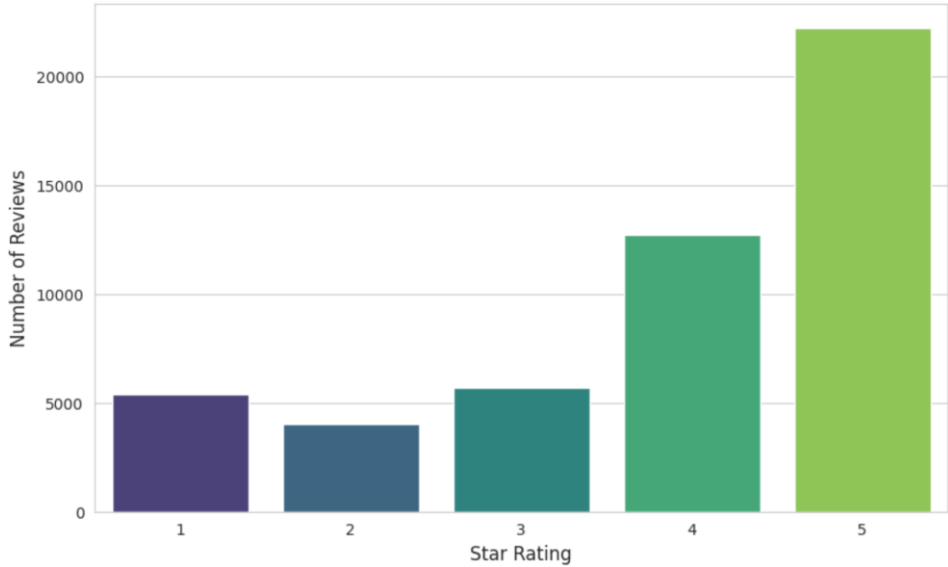


Figure 4.1 Distribution of Yelp Review Star Ratings

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.

CHAPTER 5

CONCLUSION AND FUTURE WORKS

5.1 Introduction

This chapter provides a systematic summary of the research project and offers a perspective on future work. It begins by revisiting the core challenges this study aims to address and outlines the key expected conclusions based on the proposed research design. Subsequently, the limitations of this research are objectively analyzed. Finally, drawing upon the study's findings and remaining challenges, this chapter proposes specific directions for future research in the field. The central focus of this research is the construction of an interpretable, high-accuracy, five-class sentiment prediction framework for Yelp data by integrating multi-source information, and the entirety of this chapter revolves around this core objective.

5.2 Research Summary and Conclusion

To address the prevalent challenges in Yelp user-generated content—namely "semantic-label bias," the opaque decision-making processes of advanced models, and the underutilization of structured metadata—this study proposes a comprehensive sentiment analysis framework. This framework is centered on the effective fusion of deep textual semantics, extracted via the pre-trained language model BERT, with structured business information such as categories and locations. Furthermore, it incorporates the SHAP (SHapley Additive exPlanations) method to bring transparency to the model's decision-making process.

Based on the research design detailed in the preceding chapters, this study is expected to yield the following core conclusions:

1. **Validation of Multi-Source Information Fusion (Answering RQ1):** It is anticipated that the classification models (e.g., Logistic Regression, Random Forest) constructed by concatenating BERT-derived semantic features with structured metadata will significantly outperform baseline models that rely solely on textual features. This expected outcome would demonstrate that structured data provides crucial contextual cues, thereby enhancing the model's ability to discriminate between complex and ambiguous sentiments.
2. **Achievement of Model Interpretability (Answering RQ2):** Through the application of the SHAP analysis tool, this study expects to successfully demystify the model's decision-making mechanisms at both global and local scales. Globally, it should identify the key features most influential to the model's predictions. Locally, it should provide clear attribution for any individual sample's prediction. This process effectively opens the model's "black box," substantially enhancing its credibility and transparency.
3. **In-depth Insight into Misclassification Mechanisms (Answering RQ3):** By combining confusion matrix analysis with SHAP, the framework is expected to systematically identify and explain the model's misclassification patterns for fine-grained ratings, particularly between easily confused adjacent classes like 4-star and 5-star reviews. This analysis will not only reveal the model's inherent weaknesses but also offer data-driven, actionable insights for subsequent iterations and optimizations.

In summary, this project puts forward a sentiment analysis framework that successfully balances predictive performance with interpretability. It not only charts an effective course for improving the accuracy of sentiment classification but, more importantly, provides a powerful methodology and toolset for understanding and trusting the outputs of complex AI models.

5.3 Research Limitations

Although this study proposes a well-defined framework, several limitations should be acknowledged:

1. **Data Scope Limitation:** The scope of this research is confined to English-language review data from the Yelp platform, excluding other languages and multimodal data sources such as images. This may limit the universal applicability of the model's conclusions.
2. **Model and Fusion Method Limitation:** To validate the core framework, this study employed established machine learning classifiers with a straightforward feature concatenation method. More advanced fusion techniques (e.g., attention mechanisms) and novel end-to-end deep learning architectures were not implemented in the current stage.
3. **Explainability Tool Limitation:** The research primarily utilizes SHAP as its core explainability tool. It does not include a comparative analysis with other methods (such as LIME) or delve into more advanced techniques like causal inference.

5.4 Future Works

Based on the findings and limitations of this study, future research could proceed in several promising directions to enhance the model's performance and practical value:

1. **Advanced Fusion Methods:** A primary direction for future work is the implementation and evaluation of more sophisticated fusion strategies. For instance, an **Attention Mechanism** could be introduced to allow the model to dynamically weigh the importance of text versus metadata features. Furthermore, exploring **Graph Neural Networks (GNNs)** to model the

complex relationships between reviews, users, and businesses could capture higher-order interactions.

2. **Expanded Data Sources and Types:** Future models could be enhanced by integrating additional metadata dimensions, such as user demographics, historical activity, or the temporal dynamics of reviews. Extending the framework to handle multilingual and multimodal (e.g., images within reviews) data also presents a valuable avenue for research.
3. **Model Generalization and Transferability:** The proposed framework could be applied to datasets from other domains, such as Amazon product reviews or social media platforms like Twitter, to validate its cross-domain and cross-platform generalizability and practical utility.
4. **Real-Time Sentiment Analysis:** Developing a real-time or streaming version of the framework would enable enterprises and stakeholders to receive immediate public opinion feedback and business insights, supporting agile decision-making in fast-paced environments.
5. **Enhanced Interpretability and Feedback Loop:** The explanations generated by SHAP could be further refined, or integrated with other XAI methods, to be made more accessible to non-technical users. A more advanced step would be to use these interpretability insights to guide the model's retraining process, creating an "interpretation-diagnosis-optimization" feedback loop to mitigate model bias and continuously improve robustness.

REFERENCES

- Abo, M. E. M., Idris, N., Mahmud, R., Qazi, A., Hashem, I. A. T., Maitama, J. Z., ... & Yang, S. (2021). A multi-criteria approach for arabic dialect sentiment analysis for online reviews: Exploiting optimal machine learning algorithm selection. *Sustainability*, 13(18), 10018.
- Alamoudi, E. S., & Alghamdi, N. S. (2021). Sentiment classification and aspect-based sentiment analysis on yelp reviews using deep learning and word embeddings. *Journal of Decision Systems*, 30(2-3), 259-281.
- Almansour, A., Alotaibi, R., & Alharbi, H. (2022). Text-rating review discrepancy (TRRD): an integrative review and implications for research. *Future Business Journal*, 8(1), 3.
- Arya, V., Mishra, A. K. M., & González Briones, A. (2022). Analysis of sentiments on the onset of COVID-19 using machine learning techniques.
- Aslam, N., Rustam, F., Lee, E., Washington, P. B., & Ashraf, I. (2022). Sentiment analysis and emotion detection on cryptocurrency related tweets using ensemble LSTM-GRU model. *Ieee Access*, 10, 39313-39324.
- Benarafa, H., Benkhalifa, M., & Akhloufi, M. (2024). An Improved SVM Noise Tolerance for Implicit Aspect Identification in Sentiment Analysis. *Journal of Advances in Information Technology*, 15(7).
- Brandão, J. G., Junior, A. P. C., Pacheco, V. M. G., Rodrigues, C. G., Belo, O. M. O., Coimbra, A. P., & Calixto, W. P. (2025). Optimization of machine learning models for sentiment analysis in social media. *Information Sciences*, 694, 121704.
- Chatzimina, M. E., Papadaki, H. A., Pontikoglou, C., & Tsiknakis, M. (2024). A comparative sentiment analysis of Greek clinical conversations using BERT, RoBERTa, GPT-2, and XLNet. *Bioengineering*, 11(6), 521.
- Dandash, M., & Asadpour, M. (2025). Personality analysis for social media users using Arabic language and its effect on sentiment analysis. *Social Network Analysis and Mining*, 15(1), 6.

- Das, R. K., Islam, M., Hasan, M. M., Razia, S., Hassan, M., & Khushbu, S. A. (2023). Sentiment analysis in multilingual context: Comparative analysis of machine learning and hybrid deep learning models. *Heliyon*, 9(9).
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)* (pp. 4171-4186).
- Fan, X., & Zhang, Z. (2024, July). A fine-grained sentiment analysis model based on multi-task learning. In *2024 4th International Symposium on Computer Technology and Information Science (ISCTIS)* (pp. 157-161). IEEE.
- Fransisca, D., Sulistyowati, I., Budi, I., Santoso, A. B., & Putra, P. K. (2021, November). Sentiment Analysis of Office Automation Application in One of Indonesian Ministries. In *2021 5th International Conference on Informatics and Computational Sciences (ICICoS)* (pp. 181-186). IEEE.
- Ghasemi, R., & Momtazi, S. (2023). How a Deep Contextualized Representation and Attention Mechanism Justifies Explainable Cross-Lingual Sentiment Analysis. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(11), 1-15.
- Ghatora, P. S., Hosseini, S. E., Pervez, S., Iqbal, M. J., & Shaukat, N. (2024). Sentiment Analysis of Product Reviews Using Machine Learning and Pre-Trained LLM. *Big Data and Cognitive Computing*, 8(12), 199.
- Golubev, A. A., & Loukachevitch, N. V. (2021). Use of bert neural network models for sentiment analysis in Russian. *Automatic Documentation and Mathematical Linguistics*, 55, 17-25.
- Han, K. X., Chien, W., Chiu, C. C., & Cheng, Y. T. (2020). Application of support vector machine (SVM) in the sentiment analysis of twitter dataset. *Applied Sciences*, 10(3), 1125.
- He, L. (2024). Enhanced twitter sentiment analysis with dual joint classifier integrating RoBERTa and BERT architectures. *Frontiers in Physics*, 12, 1477714.
- He, R., & McAuley, J. (2016, April). Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *Proceedings of the 25th international conference on world wide web* (pp. 507-517).

- Jain, R., Rai, R. S., Jain, S., Ahluwalia, R., & Gupta, J. (2023). Real time sentiment analysis of natural language using multimedia input. *Multimedia Tools and Applications*, 82(26), 41021-41036.
- Katić, T., & Milićević, N. (2018, September). Comparing sentiment analysis and document representation methods of amazon reviews. In *2018 IEEE 16th international symposium on intelligent systems and informatics (SISY)* (pp. 000283-000286). IEEE.
- Kaur, P. (2022). Sentiment analysis using web scraping for live news data with machine learning algorithms. *Materials today: proceedings*, 65, 3333-3341.
- Khan, M. N., Khan, M. J., & Kashif, S. (2021). The Role of User Generated Content in Shaping a Business's Reputation on Social Media: Moderating role of trust propensity. *International Journal of Marketing, Communication and New Media*, 9(16).
- Lak, A. J., Boostani, R., Alenizi, F. A., Mohammed, A. S., & Fakhrahmad, S. M. (2024). RoBERTa, ResNeXt and BiLSTM with self-attention: The ultimate trio for customer sentiment analysis. *Applied Soft Computing*, 164, 112018.
- Lak, P., & Turetken, O. (2014, January). Star ratings versus sentiment analysis--a comparison of explicit and implicit measures of opinions. In *2014 47th Hawaii international conference on system sciences* (pp. 796-805). IEEE.
- Lovera, F. A., Cardinale, Y. C., & Homsı, M. N. (2021). Sentiment analysis in Twitter based on knowledge graph and deep learning classification. *Electronics*, 10(22), 2739.
- Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30.
- Maas, A., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011, June). Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies* (pp. 142-150).
- Nursal, A. T., Omar, M. F., Nawı, M. N. M., Khalid, M. S., Hanafi, M. H., & Deraman, R. (2025). Battle of Sentiment Lexicons: Wordnet, Sentiwordnet, Textblob and Vader in Web Forum Analysis. *Journal of Information Systems Engineering and Management*, 10(2s).

- Padhy, M., Modibbo, U. M., Rautray, R., Tripathy, S. S., & Bebortta, S. (2024). Application of Machine Learning Techniques to Classify Twitter Sentiments Using Vectorization Techniques. *Algorithms*, 17(11), 486.
- Qi, Y., & Shabrina, Z. (2023). Sentiment analysis using Twitter data: a comparative application of lexicon-and machine-learning-based approach. *Social network analysis and mining*, 13(1), 31.
- Rana, M. R. R., Nawaz, A., Ali, T., Alattas, A. S., & Abdelminaam, D. S. (2024). Sentiment Analysis of Product Reviews Using Transformer Enhanced 1D-CNN and BiLSTM. *Cybernetics and Information Technologies*, 24(3), 112-131.
- Ravi, K., & Ravi, V. (2015). A survey on opinion mining and sentiment analysis: tasks, approaches and applications. *Knowledge-based systems*, 89, 14-46.
- Rizinski, M., Peshov, H., Mishev, K., Jovanovik, M., & Trajanov, D. (2024). Sentiment analysis in finance: From transformers back to explainable lexicons (xlex). *IEEE Access*, 12, 7170-7198.
- 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.
- Rogers, A., Kovaleva, O., & Rumshisky, A. (2021). A primer in BERTology: What we know about how BERT works. *Transactions of the association for computational linguistics*, 8, 842-866.
- Saudy, R. E., Alaa El Din, M., Nasr, E. S., & Gheith, M. H. (2022). A novel hybrid sentiment analysis classification approach for mobile applications Arabic slang reviews. *International Journal of Advanced Computer Science and Applications*, 13(8).
- Sharma, N. A., Ali, A. S., & Kabir, M. A. (2024). A review of sentiment analysis: tasks, applications, and deep learning techniques. *International journal of data science and analytics*, 1-38.
- Singh, A., Kalra, N., Singh, A., & Sharma, S. (2022, March). Sentiment analysis of Twitter data during Farmers' Protest in India through Machine Learning. In *2022 International Conference on Computer Science and Software Engineering (CSASE)* (pp. 121-126). IEEE.

- Singh, N., & Jaiswal, U. C. (2023). Sentiment analysis using machine learning: A comparative study. *ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal*, 12, e26785-e26785.
- Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2), 267-307.
- Tutek, M., & Šnajder, J. (2022). Toward practical usage of the attention mechanism as a tool for interpretability. *IEEE access*, 10, 47011-47030.
- Wang, W. (2023). *Sentiment Analysis: A Systematic Case Study with Yelp Scores*. *Advances in Artificial Intelligence and Machine Learning*. 2023; 3 (3): 74. Machine Learning.
- Wang, Y., Guo, J., Yuan, C., & Li, B. (2022). Sentiment analysis of Twitter data. *Applied Sciences*, 12(22), 11775.
- Wu, O., Yang, T., Li, M., & Li, M. (2020). Two-level LSTM for sentiment analysis with lexicon embedding and polar flipping. *IEEE Transactions on Cybernetics*, 52(5), 3867-3879.
- Xu, Y., Wu, X., & Wang, Q. (2015, December). Sentiment analysis of yelp's ratings based on text reviews. In *2015 17th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC)* (Vol. 17, No. 1, pp. 117-120).
- Zhao, H., Yao, Q., Song, Y., Kwok, J. T., & Lee, D. L. (2021). Side information fusion for recommender systems over heterogeneous information network. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 15(4), 1-32.