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 Background of the Problem

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.

REFERENCES

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).

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, *30*.

Qu, Y., Li, F., Li, L., Dou, X., & Wang, H. (2022). Can we predict student performance based on tabular and textual data?. *IEEE Access*, *10*, 86008-86019.

Ravi, K., & Ravi, V. (2015). A survey on opinion mining and sentiment analysis: tasks, approaches and applications. *Knowledge-based systems*, *89*, 14-46.

Rodríguez-Ibánez, M., Casánez-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.

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.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational linguistics*, *37*(2), 267-307.

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.