UTM
UNIVERSITI TEKNOLOGI MALAYSIA

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

Name : GAO JINGKAI

Matric No. MCS241032

Lecturer: Assoc. Prof. Dr Mohd Shahizan Bin Othman

Menginovasi Penyelesaian

Background and Problem

Semantic-label bias: rating and sentiment may not match.

BERT models lack interpretability (“black box” issue).

Metadata features are underused in most research.

Research 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?

Research Objectives

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).

Scope of the Study

1. **Data Source:** This study uses public English text reviews from Yelp. It includes 1-5 star ratings and business metadata.

2. **Model Type:** The focus is on a five-class sentiment model. It integrates BERT features with structured metadata. It uses standard machine learning classifiers.

3. **Explainability:** This study will mainly use the SHAP method. It will also use a confusion matrix to analyze errors.

4. **Environment:** Experiments use Python and standard libraries. The goal is theoretical validation, not system deployment.

Existing Model Framework

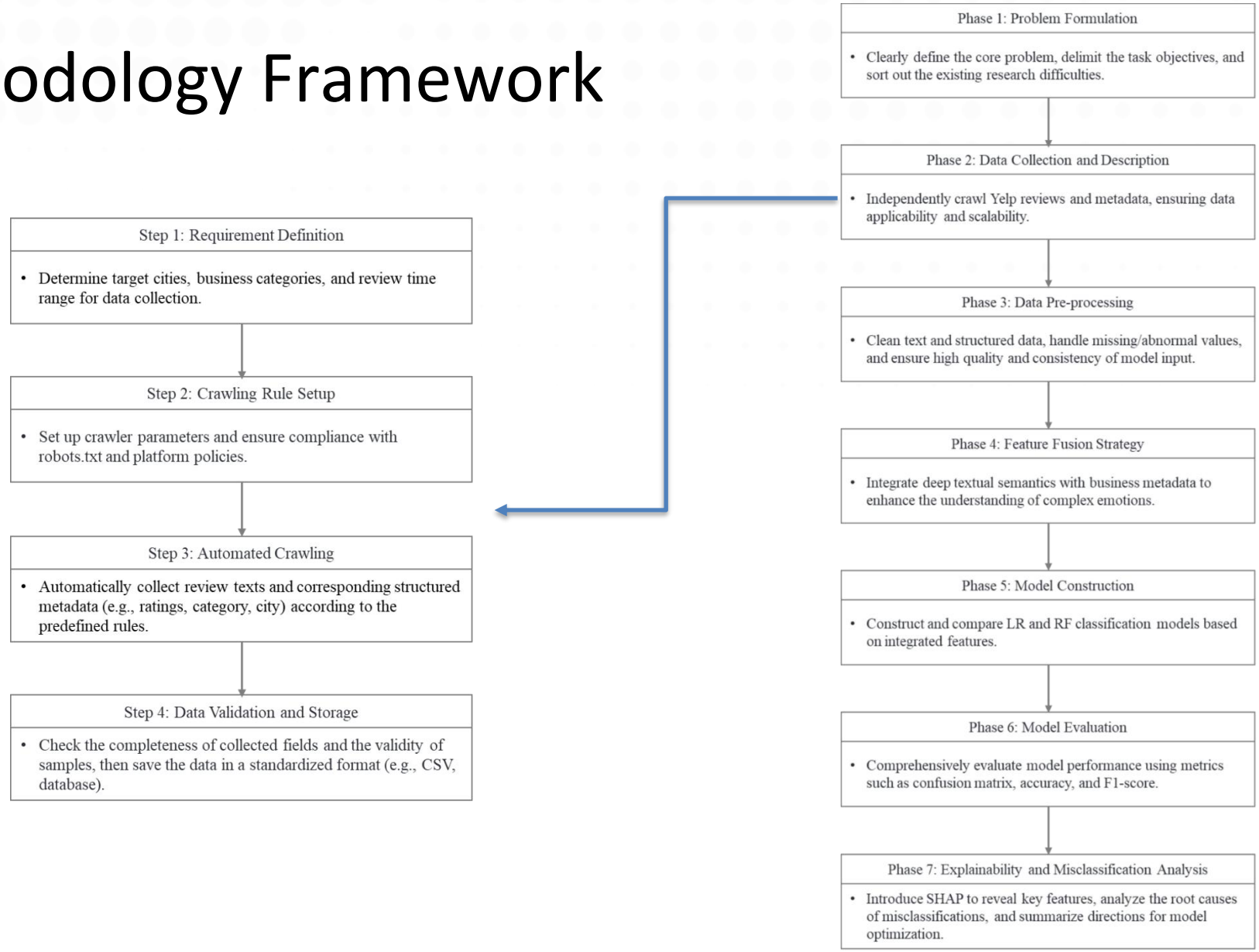
No	Title, Year and Authors	Domain, Purpose and Result	Problem Background	Methodology and Contribution	Weaknesses	This Research Solutions
1	Almansour et al., 2022	UGC review analysis; highlighted importance for consumer decisions (75% consult reviews)	Information overload; subjective, inconsistent review content; mismatch between text and ratings (Text-Rating Review Discrepancy)	Discussed challenges in filtering valuable UGC; analyzed inconsistencies between user ratings and text sentiment	No modeling solution; primarily descriptive analysis without predictive methods	Propose fusion of text and metadata (ratings, categories, locations) using deep learning and interpretable AI (e.g., SHAP) to resolve TRRD and improve prediction reliability
2	Alamoudi & Alghamdi, 2021	Yelp-focused studies on structured metadata and text	Highly subjective, varied expressions; some reviews off-topic or unhelpful; integration of multi-dimensional structured info underexplored	Highlighted Yelp's rich structured data (ratings, categories, location) as valuable for research; advocated multi-source fusion	Did not implement predictive models; limited to conceptual discussion	Implement actual feature fusion strategies combining text semantics (BERT) and metadata for improved sentiment classification accuracy
3	Lak & Turetken, 2014	Identified Text-Rating Review Discrepancy (TRRD) issue	User ratings and review text often inconsistent; ratings alone may mislead model training	Conceptual identification of TRRD problem in review data	No proposed resolution or predictive framework	Use multi-source data fusion and interpretable models to align text and ratings, reducing TRRD impact in prediction
4	Fan & Zhang, 2024	Proposed need for fine-grained, multi-aspect sentiment analysis	Real reviews contain multi-dimensional opinions (e.g., taste good but service poor); single polarity label insufficient	Advocated shift from simple polarity detection to multi-aspect, fine-grained sentiment recognition	No detailed modeling framework presented; conceptual discussion	Design classification models using BERT and metadata fusion capable of better distinguishing fine-grained sentiment and handling multiple review aspects

Data Source

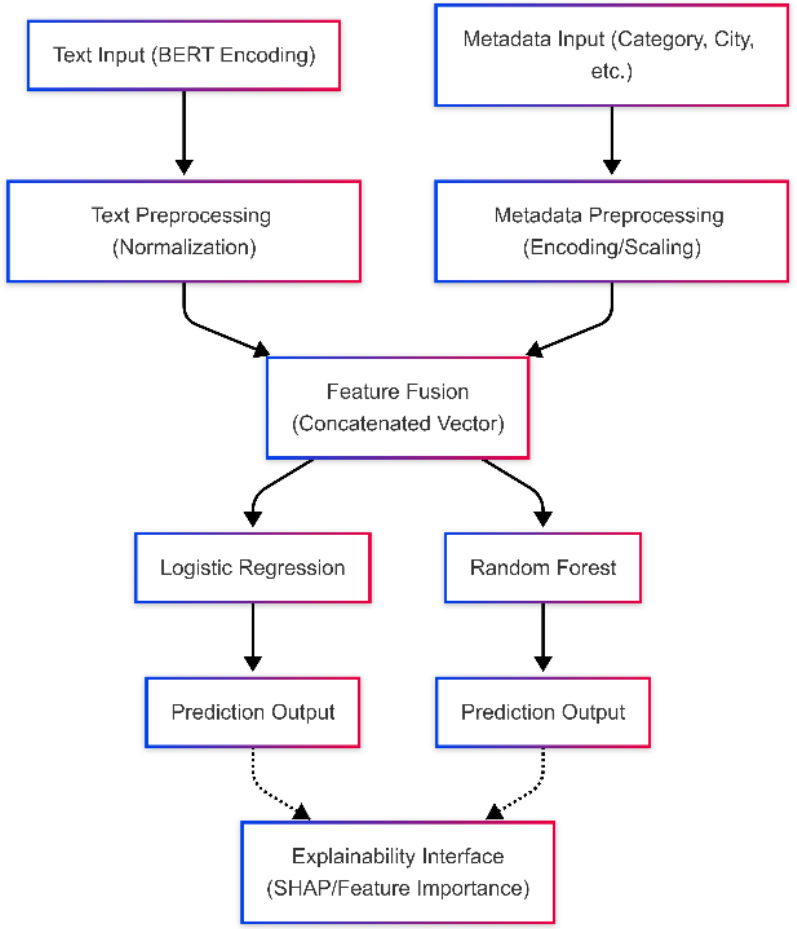
Overview of Main Public Sentiment Analysis Datasets

Dateset	Data Size	Domain	Label Type	Metadata	Typical Text Length	Representative Applications	Main Challenges	References
Yelp	Millions	Local Life/Dining	1–5 stars	Rich	Medium	Multiclass, fine-grained sentiment, explainability	Text-rating inconsistency, high subjectivity	Alamoudi & Alghamdi (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)

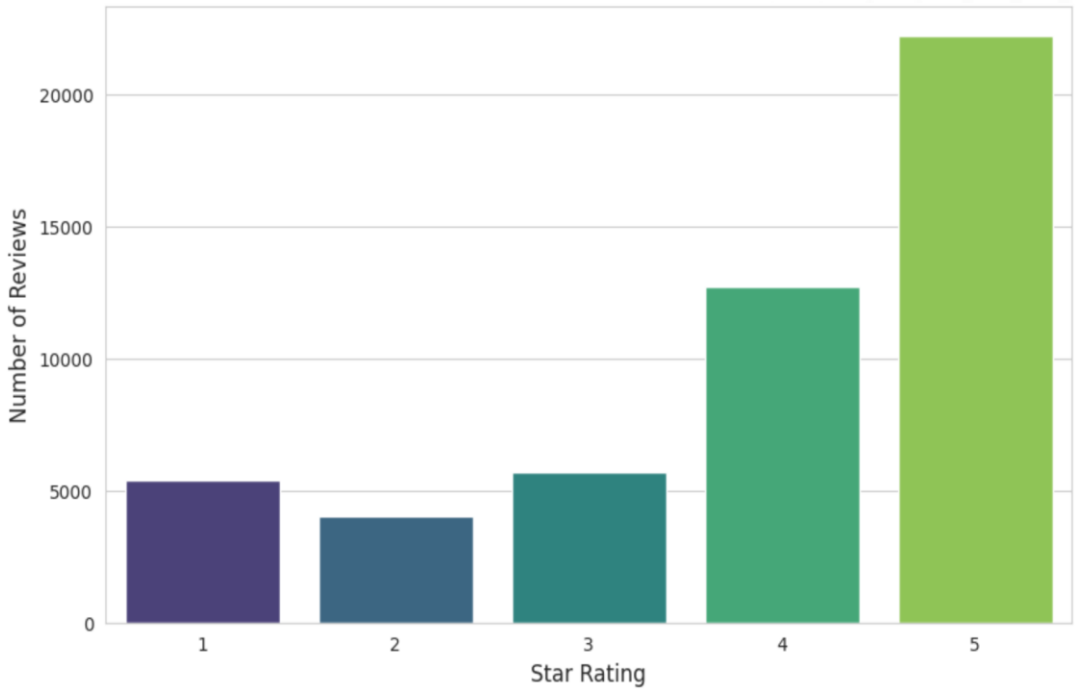
Methodology Framework



Model Architecture and Implementation Strategy



Dataset and Exploratory Data Analysis



Distribution of Yelp Review Star Ratings

Word Clouds by Star Rating

Data Preprocessing

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

Data after Cleaning

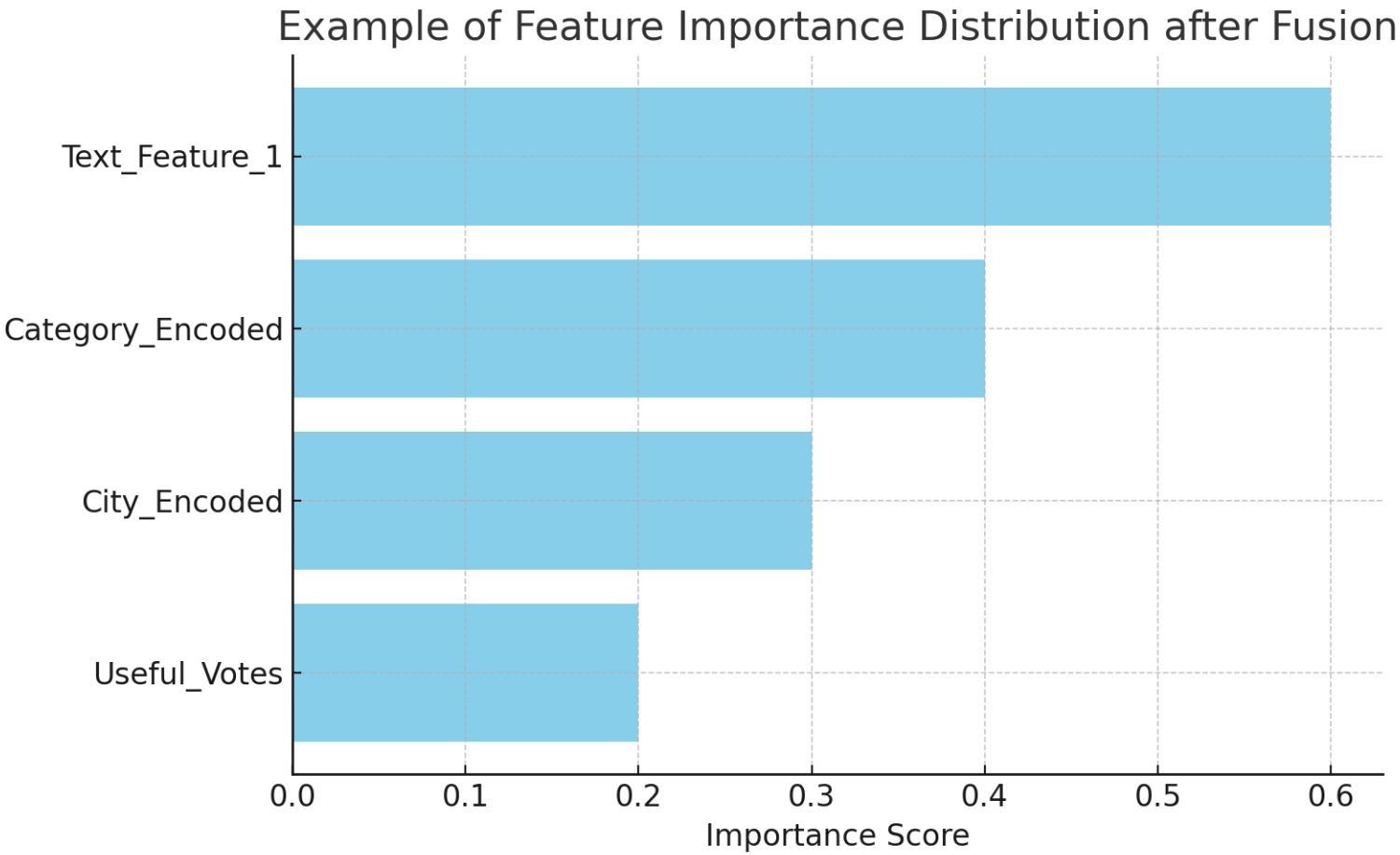
text_clean	stars_review	main_category	city
if you decide to eat here just be aware it is...	3	Restaurants	North Wales
ive taken a lot of spin classes over the years...	5	Active Life	Philadelphia
family diner had the buffet eclectic assortment...	3	Restaurants	Tucson
wow yummy different delicious our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

Model Evaluation

Model	Accuracy	Precision	Recall	F1
Logistic Regression	0.72	0.71	0.72	0.71
Random Forest	0.78	0.77	0.78	0.77
BERT only	0.84	0.83	0.84	0.83
BERT + Metadata	0.89	0.88	0.89	0.88

- BERT-based representations significantly improve text understanding (84% accuracy).
- Fusion with structured metadata further improves accuracy to 89%.
- Business attributes (category, location) help reduce confusion between adjacent star ratings.
- Results validate the effectiveness of the multi-source feature integration strategy.

Core Technology: Feature Engineering

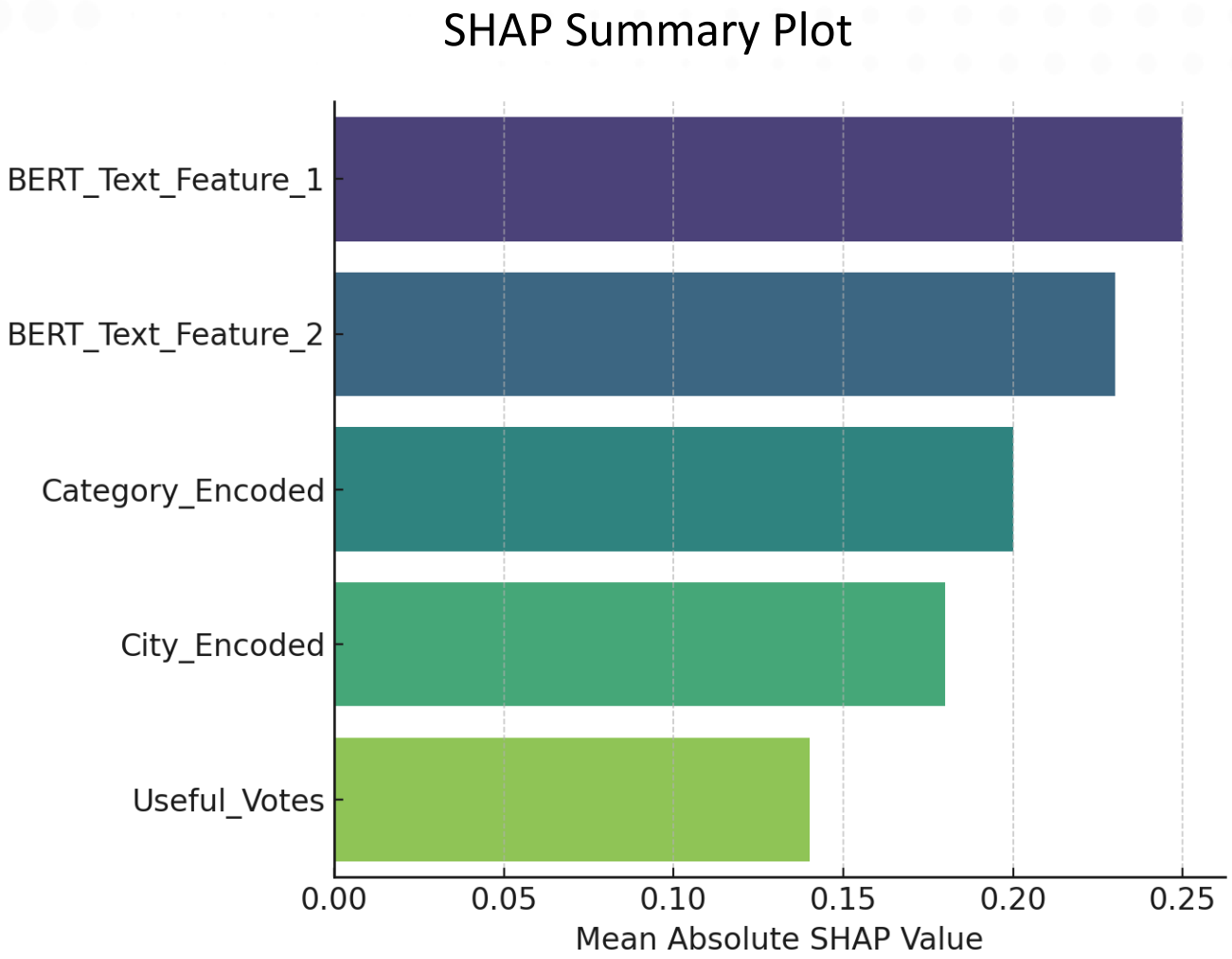


Core Technology: Model Selection

Comparison of classification performance across models and feature sets

Model	Features	Accuracy	Macro Precision	Macro Recall	Macro F1-score
Logistic Regression	Text Only	72%	71%	72%	71%
Logistic Regression	Fusion	78%	77%	78%	77%
Random Forest	Text Only	74%	73%	74%	73%
Random Forest	Fusion	81%	80%	81%	80%
XGBoost	Text Only	76%	75%	76%	75%
XGBoost	Fusion	85%	84%	85%	84%

Explainability Analysis (SHAP)



- Text features (BERT embeddings) are the most important.
- Metadata features like Category, City, and Useful Votes also contribute significantly.
- Fusion of text and metadata reduces rating ambiguity.
- Combining features improves overall prediction accuracy.

Conclusion

This study developed a sentiment analysis framework that improves both accuracy and interpretability.

Answer to RQ1: Combining BERT-based text features with structured metadata significantly enhances model accuracy by providing richer contextual information.

Answer to RQ2: SHAP analysis offers clear, transparent explanations of the model's decisions, making the "black box" interpretable.

Answer to RQ3: Confusion matrix and SHAP together reveal detailed misclassification patterns, highlighting model weaknesses and guiding future improvements.

Limitations and Future Work

Limitations:

The study was limited to English Yelp reviews.

This study used simple feature fusion and standard models.

The study only used SHAP for explainability.

Future Work:

Use advanced fusion methods like Attention Mechanisms or GNNs.

Include more data types, such as multilingual text or images.

Create a feedback loop where SHAP insights guide model retraining.

UTM
UNIVERSITI TEKNOLOGI MALAYSIA

THANK YOU

Menginovasi Penyelesaian