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

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.

Literature Review

1. Pre-trained models like BERT perform well in sentiment analysis. However, they are often "black box" models.

2. Explainability methods like SHAP are important. They can help understand model decisions.

3. This review identified three main research gaps:
   - **Misclassification:** The problem of confusing similar ratings (e.g., 4-star vs 5-star) is not well studied.
   - **Explainability Integration:** Tools like SHAP are not well integrated with misclassification analysis.
   - **Metadata Use:** Most studies ignore structured metadata in explainable models.

Methodology Framework

Phase 1: Problem Formulation

Phase 2: Data Collection and Description

Phase 3: Data Pre-processing

Phase 4: Feature Fusion Strategy

Phase 5: Model Construction

Phase 6: Model Evaluation

Phase 7: Explainability and Misclassification Analysis

Core Technology: Feature Engineering

1. **Text Features:** This study uses a pre-trained BERT model to encode text. BERT captures deep semantic information in reviews. It creates a high-dimensional vector for each review.

2. **Metadata Features:** This study uses structured data like business category and city. This data is converted into numerical vectors using One-Hot Encoding.

3. **Feature Fusion:** The study combines the text and metadata vectors. This is done through simple vector concatenation. This creates a single, unified feature vector for the model.

Core Technology: Model Selection

Classification Models: This study uses three main models:

1. **Logistic Regression (LR):** It is efficient and interpretable. It works well for high-dimensional data.
2. **Random Forest (RF):** It can capture complex non-linear patterns. It also helps analyze feature importance.
3. **XGBoost:** It is a powerful and efficient gradient boosting model used for baseline comparison.

Explainability Tool:

- This study uses **SHAP** (SHapley Additive exPlanations).
- SHAP explains how each feature contributes to a prediction. It provides both global and local explanations.

Misclassification Analysis

Confusion Matrix: A heatmap of the confusion matrix was used to find common errors. The model often confuses adjacent ratings, like 4-star and 5-star reviews.

SHAP Global Importance: SHAP summary plots were used to find the most important features overall. This shows if text or metadata has more influence.

SHAP Local Explanation: Specific misclassified reviews were chosen. SHAP force plots were used to see why the model made a mistake.

Conclusion

This study developed a sentiment analysis framework that is both accurate and interpretable.

Answer to RQ1: Fusing text features with metadata improves model accuracy.

Answer to RQ2: SHAP successfully explained the model's decisions. It opened the "black box.".

Answer to RQ3: The study analyzed misclassification patterns using the confusion matrix and SHAP. This gives insight into model weaknesses.

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.

Thank You