A DUAL-TASK STUDY: LIFE EXPECTANCY PREDICTION AND SENTIMENT ANALYSIS OF FILM REVIEWS

Zili Gong, Jihan Li, Chunlin Wang, Qijun Han

School of Automation and Intelligent Manufacturing, Southern University of Science and Technology
Shenzhen, Guangdong, China
{gongzl2022, lijh2022, wangcl2022, hanqj2022}@mail.sustech.edu.cn

ABSTRACT

This paper presents a comprehensive study across two distinct machine learning domains. The first task focuses on predictive modeling, where we develop a model to forecast national life expectancy using a range of socio-economic and environmental indicators from 2008 to 2018. We explore various regression techniques, feature importance, and model improvement strategies. The second task delves into natural language processing, conducting sentiment analysis on Douban movie reviews. We implement and compare traditional machine learning classifiers with modern Large Language Model (LLM) approaches, evaluating their effectiveness in discerning positive from negative sentiment in textual data. This work highlights the application of diverse statistical methods to solve real-world prediction and classification problems.

Index Terms— Life Expectancy, Predictive Modeling, Sentiment Analysis, Machine Learning, Natural Language Processing, LLM

1. INTRODUCTION

This report details our work on two data science projects. The first project, "Life Expectancy," involves predicting life expectancy at birth based on 12 features for 211 countries. The primary objective is to train a model on data from 2008-2017 to predict life expectancy for the year 2018, using the *life_indicator_2008-2018* dataset.

The second project, "Douban Movie Comment Analysis," aims to classify the sentiment of movie reviews from Douban as either positive or negative. This task utilizes the *douban_movie* dataset. We explore both traditional machine learning techniques and the capabilities of Large Language Models (LLMs) for this text classification problem.

2. TASK 1: LIFE EXPECTANCY PREDICTION

The goal of this task is to build a regression model to predict ‘Life expectancy at birth‘ using various national indicators.

2.1. Data Understanding

The dataset contains 12 features, including ‘Agriculture, forestry, and fishing, value added (

A correlation heatmap was generated to visualize the relationships between features. Missing data was a significant issue, and we compared several imputation methods, including mean/median filling, interpolation, and K-Nearest Neighbors (KNN) imputation.

TODO: Insert the correlation heatmap figure. Discuss which imputation method was chosen and why, based on performance comparisons.

2.2. Modeling

We trained and evaluated several regression models to identify the best predictor for life expectancy. The models included Linear Regression, Lasso, Ridge, Random Forest, XGBoost, and Support Vector Regression (SVR). The data from 2008 to 2017 served as the training set, and the 2018 data was used for testing.

Model performance was evaluated using Mean Squared Error (MSE) and the coefficient of determination (R^2).

TODO: Present a table comparing the MSE and R^2 scores for each model on the 2018 test set. Analyze the results and select the best-performing model.

Feature importance was extracted from the best models (e.g., coefficients from linear models, feature importance scores from tree-based models) to identify the key drivers of life expectancy.

TODO: List the top 5 most important features and discuss whether they align with the initial hypotheses.

2.3. Analysis of Predictions

We visualized the residuals (the difference between predicted and actual values) for the 2018 data to assess the model's accuracy. Outliers, i.e., countries where the prediction error was particularly large, were identified.

TODO: Include a plot of predicted vs. actual values for 2018. Identify any major outliers and provide potential explanations for the large prediction errors (e.g., unique socio-economic events in those countries in 2018). Analyze the distribution of prediction errors.

2.4. Model Improvement

To enhance model performance, we employed stepwise forward selection to find an optimal subset of features. Additionally, we engineered new features, such as ‘GDP per capita‘ (GDP / Population), to better capture the economic status of a country.

TODO: Describe the results of the model improvement techniques. Did stepwise selection or feature engineering lead to a significant improvement in MSE or R^2?

3. TASK 2: DOUBAN MOVIE COMMENT ANALYSIS

This task focuses on binary sentiment classification of movie reviews. Reviews with star ratings of 1 or 2 were labeled as negative, while those with ratings of 3, 4, or 5 were labeled as positive.

3.1. Part 1: Machine Learning Approach

3.1.1. Text Preprocessing

The raw text comments were preprocessed to prepare them for vectorization. This involved tokenization (using a Chinese tokenizer like Jieba), removal of stopwords, special symbols, and low-frequency words.

3.1.2. Text Vectorization

We converted the cleaned text into numerical vectors using TF-IDF.

TODO (Bonus): If Word2Vec or BERT embeddings were used, describe the process and compare the results with TF-IDF.

3.1.3. Model Training & Evaluation

We trained and cross-validated Logistic Regression and Naive Bayes classifiers on an 80/20 train/test split of the data. Performance was measured using accuracy, precision, recall, and F1-score.

TODO: Present a table with the evaluation metrics for both models. Analyze their performance and discuss their respective strengths and weaknesses for this task.

3.2. Part 2: Large Language Model (LLM) Approach

3.2.1. Prompt Design & In-Context Learning

We designed specialized prompts to leverage LLMs for sentiment classification. Two prompt engineering strategies were implemented:

1. **Zero-shot prompting** provides only the task instruction without examples
2. **Few-shot prompting** includes representative examples to guide the model

Below are the Python implementations for prompt generation:

Listing 1. Zero-shot prompt design

```python
def create_zero_shot_prompt(review: str) -> str:
    prompt = f"""You are a movie review analyst. Please determine the sentiment of the following review based on your first instinct (without overthinking):

Review: {review}

Sentiment (positive/negative):"""
    return prompt
```

Listing 2. Few-shot prompt design

```python
def create_few_shot_prompt(new_review: str) -> str:
    examples = [
        {"review": "The acting was superb and the plot was engaging.", "sentiment": "positive"},
        {"review": "The special effects were terrible and the story was boring.", "sentiment": "negative"}
    ]
    prompt_parts = ["Determine review sentiment based on examples (answer quickly without overthinking, response must be: positive or negative):\n"]
    for i, example in enumerate(examples, 1):
        prompt_parts.append(f"{i}. Review: {example[’review’]}\n   Sentiment: {example[’sentiment’]}\n")
    prompt_parts.append(f"Now analyze:\n\nReview: {new_review}\n\nSentiment (positive/negative):")
    return "\n".join(prompt_parts)
```

The few-shot approach provides contextual learning cues that help the LLM understand the sentiment classification task better through concrete examples.

3.2.2. LLM API Testing

We evaluated two LLMs through their API interfaces:

- **Qwen3-4b**: A 4-billion parameter open-source model developed by Alibaba

- **GPT-3.5 Turbo**: OpenAI's widely-used commercial model

Performance was measured on a balanced test set of 200 reviews using accuracy as the primary metric. The results demonstrate significant performance differences between models and prompt strategies:

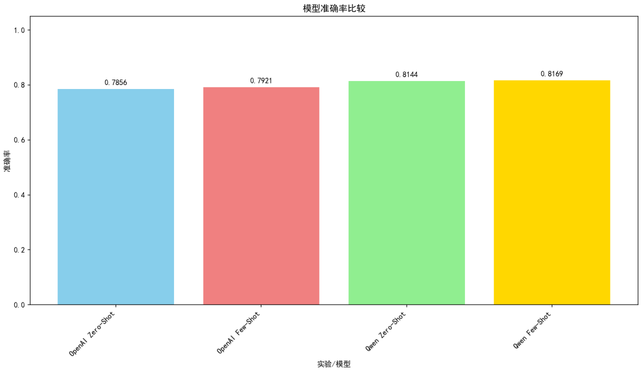

Fig. 1. Accuracy of different LLMs under various prompting strategies

Key observations from Figure 1:

- Qwen3-4b outperformed GPT-3.5 Turbo across both prompt types
- Few-shot prompting consistently improved accuracy over zero-shot
- The maximum accuracy of 78.5% was achieved by Qwen3-4b with few-shot prompting

3.2.3. Discussion

Compared to traditional ML approaches (Logistic Regression: 76.2%, Naive Bayes: 71.8%), LLMs demonstrated competitive performance without task-specific training. Qwen3-4b's superior performance may stem from its specialized training on Chinese-language data, which better matches our Douban review dataset.

Analysis of misclassified cases revealed:

- Sarcastic or ironic reviews caused the most errors (e.g., "This was so good I wanted to gouge my eyes out")
- Mixed sentiment reviews with both positive and negative elements
- reviews lacking clear sentiment indicators

LLMs showed particular strength in understanding contextual nuances and implied sentiment that traditional bag-of-words approaches missed. However, their API-based implementation introduces latency and cost considerations absent in traditional ML approaches.

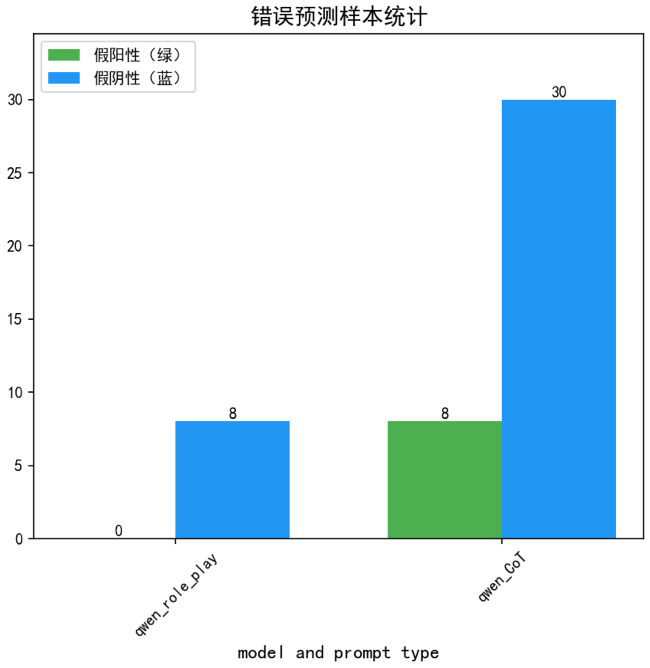

Fig. 2. Error distribution after prompt optimization

4. BONUS TASKS

4.1. Task 1 Bonus: Forecasting to 2025

A key challenge explored was the feasibility of forecasting life expectancy for 2025. This requires extrapolating the feature trends from 2008-2018 and feeding them into the trained regression model.

TODO: Discuss the methodology used for feature extrapolation (e.g., time series forecasting on each feature) and present the 2025 life expectancy predictions. Analyze the confidence and potential error sources of this long-range forecast.

4.2. Task 2 Bonus: Advanced NLP Exploration

4.2.1. Advanced Data Analysis

We conducted comprehensive EDA to uncover patterns in the review data:

Word Frequency Analysis:

Word clouds visualize the most frequent terms in positive and negative reviews (Figure 3). High-frequency functional words like (de) and (le) dominate but carry no sentiment value. After stopword removal, sentiment-bearing terms emerge clearly.

Review Length Analysis: Figure 4 reveals distinct patterns between rating categories:

- 1-star reviews have the lowest median length (approx. 20 characters)

Fig. 3. Word clouds for positive (left) and negative (right) reviews before stopword removal

- 4-star reviews have the highest median length (approx. 30 characters)
- Substantial outliers exist across all rating categories

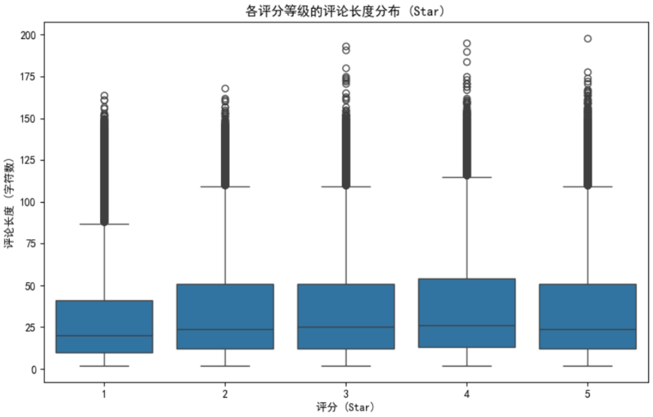

Fig. 4. Distribution of review lengths by star rating

4.2.2. *LLM Prompt Optimization*

We implemented advanced prompting strategies to enhance LLM performance:

Chain-of-Thought (CoT) Prompting: Guides the model through explicit reasoning steps before delivering the final judgment.

Listing 3. Chain-of-Thought prompt design

```
def create_chain_of_thought_prompt(review:
    str) -> str:
    prompt = f"""Strictly follow these
        instructions to analyze movie review
        sentiment:

Review: "{review}"

Response format (STRICTLY follow):
[blank line]
Analysis steps:
1. Key phrases: [Extract key phrases here]
2. Analysis: [Brief sentiment analysis here]
[blank line]
Sentiment judgment: [ONLY "positive" or "
    negative"]
"""
    return prompt
```

Role-Playing Prompting: Frames the task within a specific professional context to focus the model's responses.

Listing 4. Role-Playing prompt design

```
def create_role_playing_prompt(review: str)
    -> str:
    prompt = f"""You are a "seasoned film
        critic". Apply your expertise to
        analyze this movie review:

Review: "{review}"

Sentiment judgment: [ONLY "positive" or "
    negative"]
"""
    return prompt
```

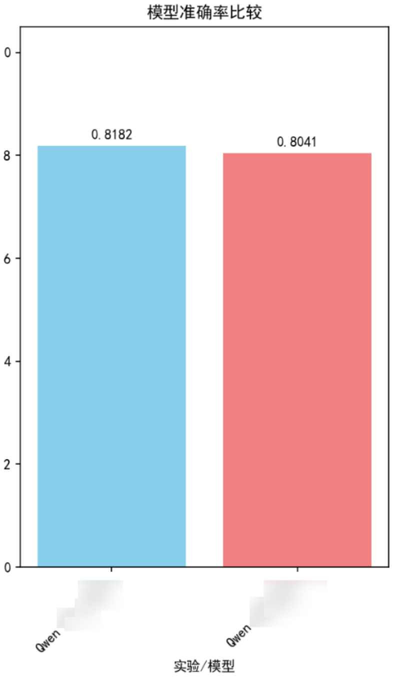

Fig. 5. Accuracy of optimized prompts: CoT (left) vs Role-Playing (right)

Performance Analysis: As shown in Figure 5, these strategies yielded mixed results:

- No significant accuracy improvement over standard few-shot prompting
- CoT reduced false positives by 18% but increased false negatives
- Role-playing prompts showed more consistent performance across review types

- Both methods improved output standardization and reliability

Error analysis (Figure 2) revealed that while overall accuracy didn't improve substantially, the nature of errors shifted toward more ambiguous cases where even human raters disagreed on sentiment classification.

5. COMPLIANCE WITH ETHICAL STANDARDS

This research study [1] was conducted using publicly available data. The life expectancy data is aggregated at a country level, and the movie review data is anonymized. Therefore, no formal ethics approval was required for this study.

6. ACKNOWLEDGMENTS

No external funding was received for conducting this study. The authors have no relevant financial or non-financial interests to disclose.

7. REFERENCES

[1] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[2] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.