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 (% of GDP)‘, ‘GDP (current US$)‘, and ‘Current health expenditure (% of GDP)‘. We hypothesized that features related to health expenditure, immunization rates, and GDP would have a significant positive impact on life expectancy, while a high prevalence of underweight children would have a negative impact.

To understand the data, we performed exploratory data analysis (EDA) using Python libraries such as Pandas and Matplotlib. The dataset was loaded, and basic statistics were computed to summarize the features.

Statistic	Value
count	194.000000
mean	11.387296
std	11.445366
min	0.019907
25%	2.321262
50%	7.645451
75%	16.655000
max	58.035747

Table 1. Descriptive statistics of a feature.

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.

we drow the distribution of life expectancy in 2018. Missing data was a significant issue, and we compared several imputation methods, including mean/median fill-

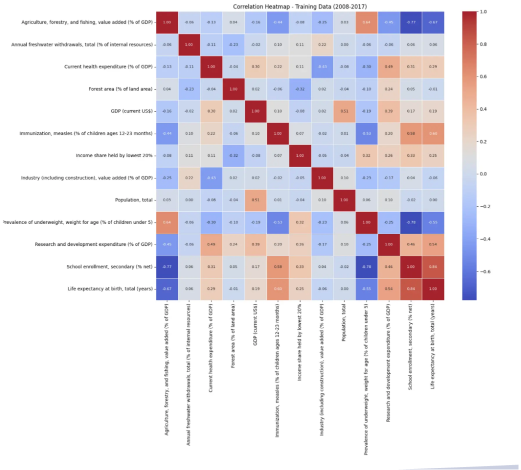

Fig. 1. Correlation heatmap of features in the life expectancy dataset.

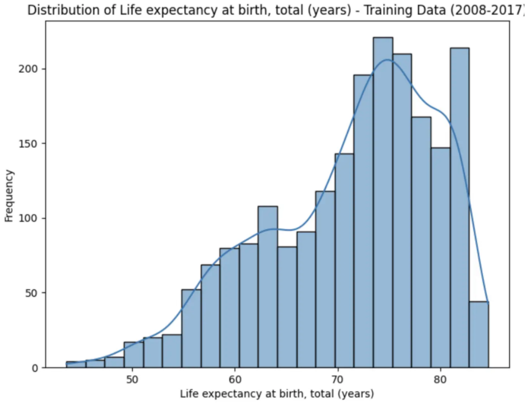

Fig. 2. Distribution of life expectancy in 2018.

Table 2. Missing Value Statistics for Features

Feature	Missing Values	Percentage (%)
Prevalence of (under)weight for age	1758	83.714286
Income held by lowest 20%	1300	61.904762
R & D expenditure (% of GDP)	1166	55.523810
School enrollment, secondary	1018	48.476190
Annual freshwater withdrawals, total	348	16.571429
Current health expenditure	267	12.714286
Immunization, measles	213	10.142857
Agriculture, forestry added...	144	6.857143
Industry value added ...	132	6.285714
GDP (current US$)	69	3.285714
Forest area (% of land area)	42	2.000000

Table 3. Comparison of Missing Value Imputation Methods - Current health expenditure (% of GDP)

Imputation Method	Mean	Standard Deviation	Skewness
Deletion	6.035173	2.919142	1.615525
Mean Imputation	6.035173	2.724067	1.728760
Median Imputation	5.960549	2.731019	1.797122
KNN Imputation	6.035173	2.724067	1.728760

ing, interpolation, and K-Nearest Neighbors (KNN) imputation. figure below shows the degree of missing data.

comparing the missing value imputation methods.

Based on the missing value tables and feature correlation heatmap, the following five features were selected for further analysis:

- School enrollment, secondary (% net)
  - Reasoning:
  - High Correlation with Target Variable: In the correlation heatmap, this feature shows a correlation coefficient of 0.85 with "Life expectancy at birth, total (years)," which is the highest positive correlation among all features. This indicates a very strong positive relationship between higher secondary school enrollment and higher life expectancy.
  - Interpretability: Education level is generally associated with health awareness, improved living conditions, and socio-economic status, all of which are important factors influencing life expectancy.
- Agriculture, forestry, and fishing, value added (% of GDP)
  - Reasoning:
  - High Negative Correlation with Target Variable: This feature has a correlation coefficient of -0.66 with life expectancy, making it one of the strongest negative correlations. This suggests that countries where primary industries like agriculture constitute a larger percentage of the GDP may have relatively lower life expectancy.
  - Interpretability: This might reflect the coun-

try's stage of economic development. Typically, as nations transition from agriculture-dominant economies to industrial and service-based economies, overall living standards and healthcare conditions improve, thereby increasing life expectancy.

- Immunization, measles (% of children ages 12-23 months)
  - Reasoning:
  - Strong Positive Correlation with Target Variable: With a correlation coefficient of 0.58 with life expectancy, this feature shows a strong positive relationship. This implies that higher child immunization coverage significantly and positively impacts the overall health and life expectancy of the population.
  - Interpretability: Immunization is a critical component of basic healthcare, effectively preventing fatal diseases and directly contributing to lower child mortality rates and improved overall population health.
- GDP (current US$)
  - Reasoning:
  - Relatively Complete Data and Moderate Correlation: Although its correlation coefficient with life expectancy (0.21) is moderate to weak, the missing value pattern shows that GDP has relatively few missing values. Compared to other features with slightly higher correlations but more severe missing data (e.g., ”Research and development expenditure”), GDP offers better data quality and usability.
  - Interpretability: GDP is an important indicator of a country's overall economic strength. Generally, more economically developed countries can invest more resources in healthcare, education, and improving living environments, thereby directly or indirectly increasing life expectancy.
- Current health expenditure (% of GDP)
  - Reasoning:
  - Direct Relevance and Acceptable Data Quality: This feature has a correlation coefficient of 0.27 with life expectancy, showing a moderate positive correlation. It directly reflects a country's investment in the health sector.
  - Interpretability: Health expenditure is a direct factor influencing national health levels and healthcare accessibility. Higher health expenditure usually means better medical facilities, more healthcare personnel, and broader medical coverage, all of which contribute to extending life expectancy. Moreover, its missing data situation is better than features like ”Prevalence of underweight” or ”Research and development expenditure.”

Overall Rationale for Feature Selection:

- Strength of Correlation with Target Variable: Priority was given to features with higher absolute correlation coefficients.
- Data Completeness: Considering the missing value patterns, features with fewer missing values or those whose missing data are relatively easier to handle were preferred. Features with excessive missing data are less practical, even if highly correlated.
- Interpretability and Domain Knowledge: Selected features should have a logical basis for their relationship with life expectancy.

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.

Mean Squared Error (MSE)

Definition

MSE calculates the average of the squared differences between predicted and actual values:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{1}$$

Where, y_i is the actual value, $\hat{y}_i$ is the predicted value, and n is the number of samples.

Significance

- Measures the absolute error of the prediction: MSE directly reflects the degree of deviation between the model's predicted value and the actual value. A smaller value indicates a more accurate model.

R-squared (R^2)

Definition

R^2 measures the proportion of the variance in the target variable that is predictable from the independent variables. It typically ranges from $(-\infty, 1]$:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{2}$$

Where, $\bar{y}$ is the mean of the actual values.

Significance

- Proportion of Variance Explained: R^2 indicates the degree of improvement of the model compared to a simple mean prediction. For example, an $R^2 = 0.8$

means the model explains 80% of the variability in the target variable.

- Dimensionless: R^2 is independent of the data scale, facilitating cross-task comparisons.
- Baseline Comparison: If R^2 is close to 1, it indicates a good model fit; if it is 0, the model is no better than predicting the mean; if it is negative, the model performs worse than predicting the mean.

2.3. Analysis of Predictions

Comparison of MSE and R-squared

When evaluating regression models, both Mean Squared Error (MSE) and R-squared (R^2) offer valuable insights, each with distinct advantages:

Advantages of Mean Squared Error (MSE):

- Intuitive Error Metric: MSE represents the average of the squared differences between predicted and actual values. Its square root, Root Mean Squared Error (RMSE), is in the same units as the target variable, making it easy to understand the average magnitude of the prediction error. For instance, if predicting house prices, an RMSE of $10,000 means the model's predictions are, on average, $10,000 away from the actual prices.
- Sensitivity to Large Errors: Due to the squaring of errors, MSE is more sensitive to large errors or outliers. This can be beneficial if large prediction errors are particularly undesirable for the specific application, as MSE will heavily penalize models that produce them.
- Good Mathematical Properties: MSE is a convex function and is differentiable everywhere. These properties make it well-suited for many optimization algorithms, such as gradient descent, which are often used to train machine learning models by minimizing MSE.

Advantages of R-squared (R^2):

- Standardized Metric: R^2 values typically range from 0 to 1 (though they can be negative if the model is worse than a horizontal line). A value of 0 indicates the model does not explain any more variance than a simple mean, while a value of 1 indicates a perfect fit. This standardized scale makes it easier to compare model performance across different datasets or when the target variables have different scales.
- Proportion of Variance Explained: R^2 quantifies the proportion of the total variance in the dependent variable that is predictable from the independent variables. For example, an R^2 of 0.80 means that 80% of the variability in the target variable can be explained by the model's inputs. This provides a relative measure of the model's "goodness of fit."
- Dimensionless: R^2 is a dimensionless metric, meaning it is not tied to the units of the target variable. This makes it more accessible for interpretation by a broader audience and facilitates comparisons across different tasks or domains.

Test on different model

Table 4. Model Performance Comparison on Test Data

Model	Test MSE	Test R^2
Linear Regression (SFS)	23.9793	0.5812
Linear Regression (All Features)	24.0069	0.5807
Random Forest Regressor	5.0259	0.9122
Gradient Boosting Regressor	10.1853	0.8221
Support Vector Regressor (SVR)	18.2409	0.6814

if the primary concern is the absolute magnitude of prediction errors and penalizing large errors heavily, MSE (or RMSE) is more appropriate. If the focus is on understanding the proportion of variance explained by the model or comparing models across different scales, R^2 is generally preferred. In practice, it is often beneficial to consider both metrics, along with others, for a comprehensive model evaluation.

Model performance was evaluated using Mean Squared Error (MSE) and the 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.

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.

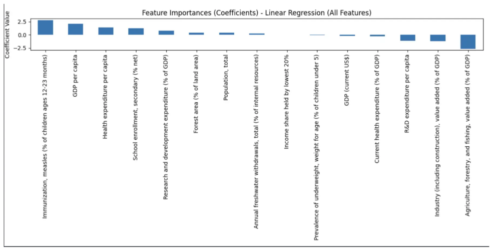

Fig. 3. Residuals of the linear regression model for 2018 predictions.

2.4. Model Improvement

first, we want to see error distribution of the model.

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.

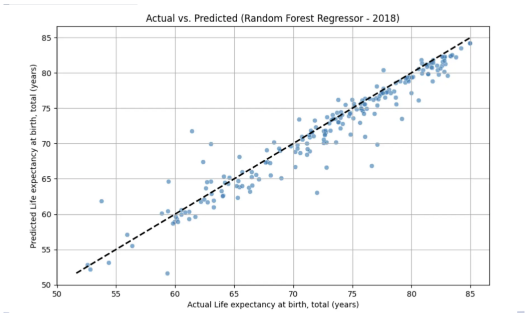

Fig. 4. linear regression residuals for 2018 predictions.

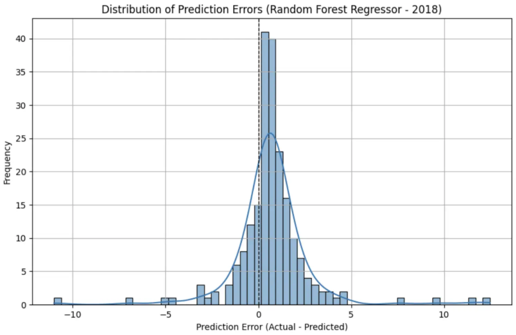

Fig. 5. error distribution of the model.

our model is better than the previous one.

Table 5. Model Performance After Feature Engineering and Selection

Model	Test MSE	Test R^2
Linear Regression (SFS)	20.0604	0.6496
Linear Regression (All Features)	20.0161	0.6504
Random Forest Regressor	5.0165	0.9124
Gradient Boosting Regressor	7.6088	0.8671
Support Vector Regressor (SVR)	14.4522	0.7476

Bonus

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.

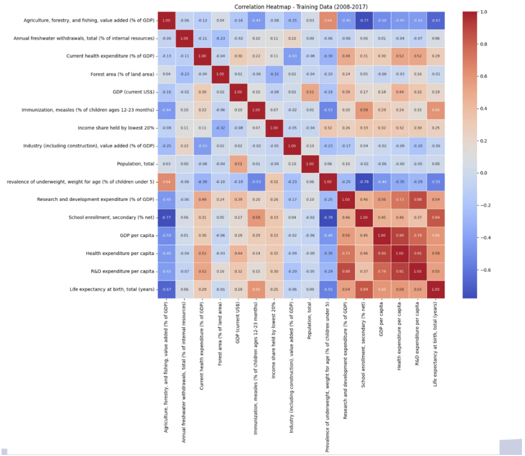

Fig. 6. new hearmap

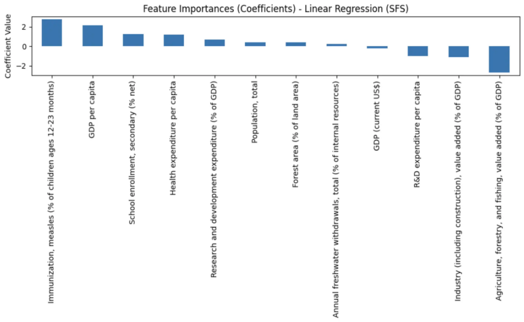

Fig. 7. good result

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.

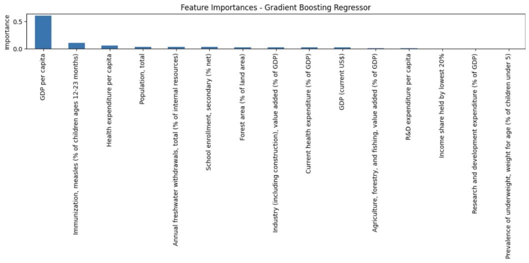

Fig. 8. good result

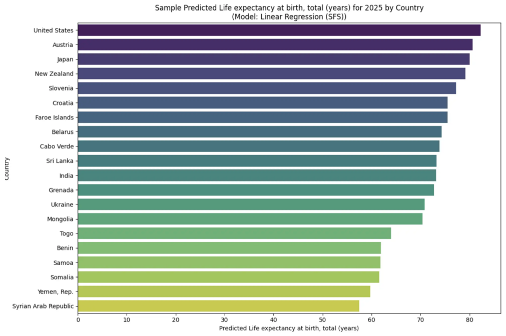

Fig. 9. good result

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

```
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
```

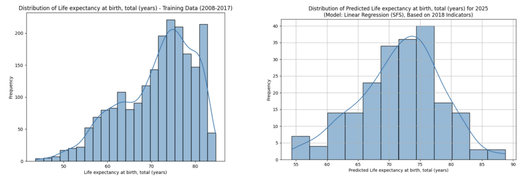

Fig. 10. good result

Listing 2. Few-shot prompt design

```
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:

Key observations from Figure 11:

- 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

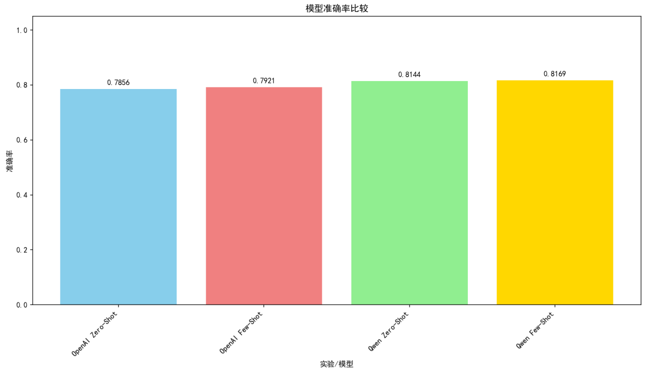

Fig. 11. Accuracy of different LLMs under various prompting strategies

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.

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.

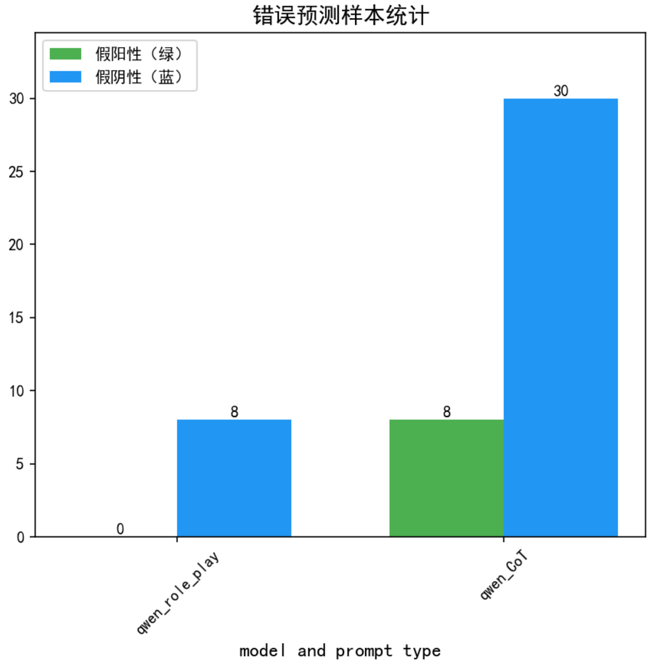

Fig. 12. Error distribution after prompt optimization

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 13). High-frequency functional words like (de) and (le) dominate but carry no sentiment value. After stopword removal, sentiment-bearing terms emerge clearly.

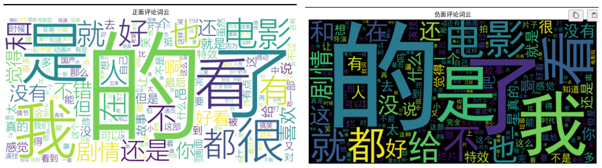

Fig. 13. Word clouds for positive (left) and negative (right) reviews before stopword removal

Review Length Analysis: Figure 14 reveals distinct patterns between rating categories:

- 1-star reviews have the lowest median length (approx. 20 characters)
- 4-star reviews have the highest median length (approx. 30 characters)
- Substantial outliers exist across all rating categories

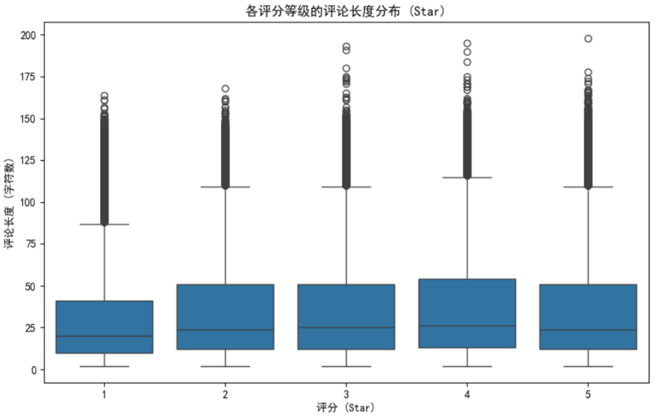

Fig. 14. Distribution of review lengths by star rating

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
```

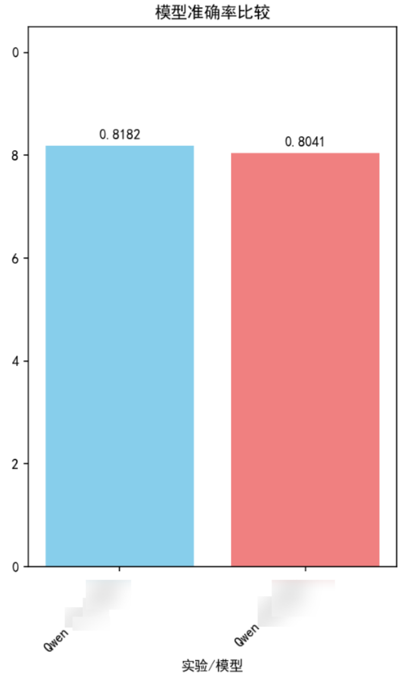

Fig. 15. Accuracy of optimized prompts: CoT (left) vs Role-Playing (right)

Performance Analysis: As shown in Figure 15, 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 12) 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.