CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

In addition, the overall research adopted in this study in the process of sentiment analysis of the movie "NEZHA 2" will be explained in detail. This study combines natural language processing (NLP) technology, covering the entire process from data collection and cleaning in the review process to classification and comparison using different sentiment analysis models. Each step has been systematically designed to ensure that the differences in emotional reactions of audiences in different countries to the film can be effectively understood in the end.

3.2 Research Framework

In order to fully complete the multi–country sentiment analysis, this study is divided into the following five stages:

– Phase 1 : Data Collection

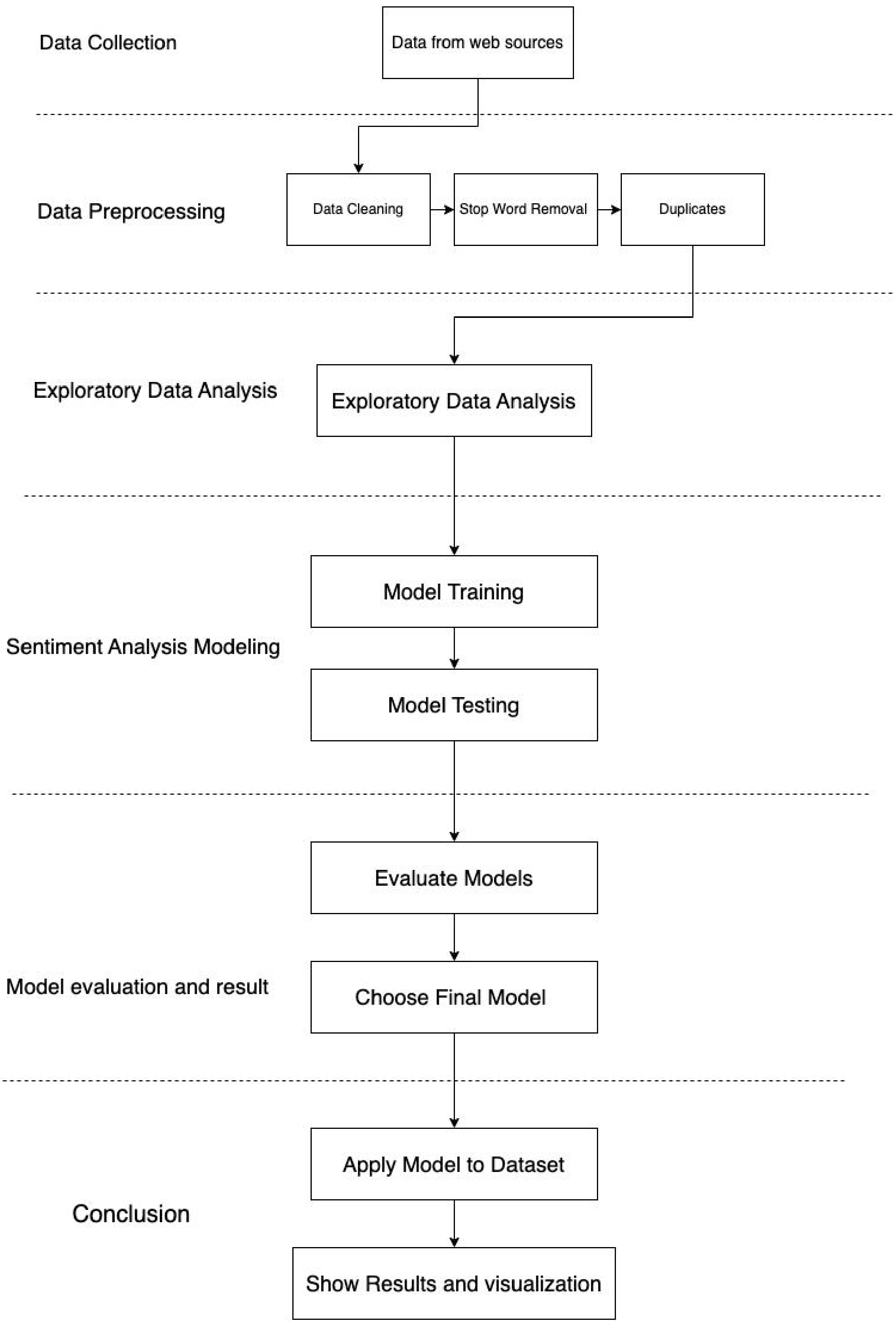
– Phase 2：Data preprocessing

– Phase 3：Exploratory Data Analysis

– Phase 4：Sentiment Analysis Modeling

– Phase 5：Model evaluation and result comparison

The following flowchart presents the methodological framework of this study.


Data Collection
Data from web sources
Data Preprocessing
Data Cleaning
Stop Word Removal
Duplicates
Exploratory Data Analysis
Exploratory Data Analysis
Model Training
Sentiment Analysis Modeling
Model Testing
Evaluate Models
Model evaluation and result
Choose Final Model
Apply Model to Dataset
Conclusion
Show Results and visualization

3.2.1 Data Collection

The main comment data comes from the following platforms IMDb

The data collected includes:

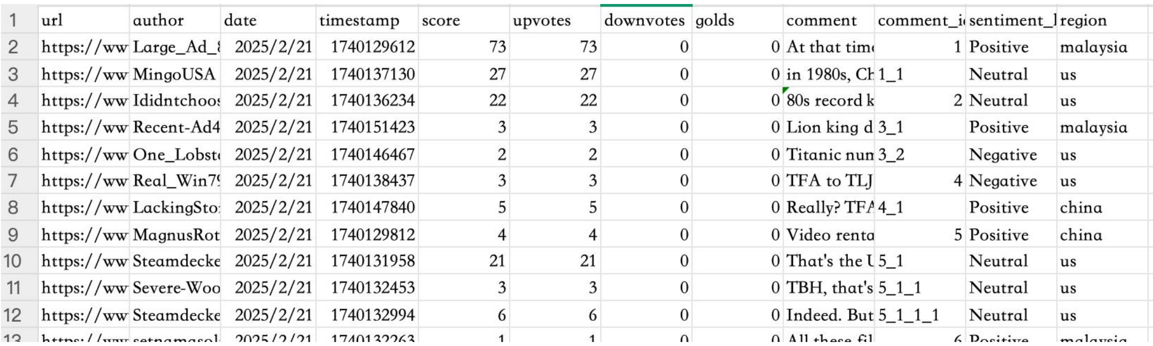
1 User comment text

2 Comment timestamp

3 Country or region information (determined by user information or IP attribution)

4 Rating (used to assist in label determination)

A total of about 20,000 reviews were collected, covering Chinese, English and some translated content, ensuring regional and linguistic diversity of the reviews.



1	url	author	date	timestamp	score	upvotes	downvotes	golds	comment	comment_i	sentiment_l	region
2	https://ww	Large_Ad_	2025/2/21	1740129612	73	73	0	0	At that tim	1	Positive	malaysia
3	https://ww	MingoUSA	2025/2/21	1740137130	27	27	0	0	in 1980s, Ch	1_1	Neutral	us
4	https://ww	Ididntchoo	2025/2/21	1740136234	22	22	0	0	80s record k	2	Neutral	us
5	https://ww	Recent-Ad4	2025/2/21	1740151423	3	3	0	0	Lion king d	3_1	Positive	malaysia
6	https://ww	One_Lobst	2025/2/21	1740146467	2	2	0	0	Titanic nun	3_2	Negative	us
7	https://ww	Real_Win7	2025/2/21	1740138437	3	3	0	0	TFA to TLJ	4	Negative	us
8	https://ww	LackingSto	2025/2/21	1740147840	5	5	0	0	Really? TF	4_1	Positive	china
9	https://ww	MagnusRot	2025/2/21	1740129812	4	4	0	0	Video renta	5	Positive	china
10	https://ww	Steamdecke	2025/2/21	1740131958	21	21	0	0	That's the U	5_1	Neutral	us
11	https://ww	Severe-Woo	2025/2/21	1740132453	3	3	0	0	TBH, that's	5_1_1	Neutral	us
12	https://ww	Steamdecke	2025/2/21	1740132994	6	6	0	0	Indeed. But	5_1_1_1	Neutral	us
13	https://ww	setnamasol	2025/2/21	1740132263	1	1	0	0	All these fil	6	Positive	malaysia

3.2.2 Data preprocessing

There is a lot of unstructured text and noise information in the original data. To ensure the quality of model input, preprocessing includes the following steps:

– Convert to lowercase

– Remove URLs, HTML tags, emoticons, special characters

– Remove stop words (Chinese and English use specific stop word lists)

– Remove empty values, duplicate or very short comments

In terms of exploratory analysis, we used word frequency statistics, word cloud generation, average score analysis and other methods to preliminarily observe the emotional tendencies and keyword differences of audiences in different countries.

3.2.3 Exploratory Data Analysis

After restoration, preliminary exploratory analysis of the data was performed to understand its basic structure and quality.

Distribution analysis of text data:

The distribution of sentiment labels is roughly balanced between positive and negative.

Text Length Distribution Most reviews are concentrated between 100 and 200 words.

Outlier data identification and processing:

Text anomalies such as being too short (less than 3 words) or too long (more than 1000 words) may not provide meaningful semantic information.

Duplicate data contains identical review text.

3.2.4 Sentiment Analysis Model

VADER: A rule–based model for English reviews, suitable for processing short texts, outputting sentiment polarity scores (–1 to +1) and classifying them into positive, neutral, and negative based on thresholds.
TextBlob: As a lightweight English sentiment analysis model, it provides polarity and subjectivity scores and is suitable as a baseline comparison model.

BERT: It is a pre–trained language model based on the Transformer architecture, which can understand the meaning of words in context and model sentence semantics through a bidirectional encoder. It is widely used in tasks such as sentiment analysis, question–answering systems, and text classification.

XGboost: It is a boosting tree algorithm that continuously stacks weak learners (usually decision trees) and optimizes the residual of the previous step at each step to make the model prediction more accurate.

Random Forest: It is an integrated algorithm that combines multiple "randomly constructed decision trees". It obtains stronger overall prediction results through "voting (classification)" or "average (regression)".

The model training set is constructed through some manually annotated reviews, and the validation set is used to evaluate the robustness of each model in different contexts.

Logistic Regression: Logistic regression is a supervised learning algorithm for binary classification problems. Its core idea is to predict the probability that a sample belongs to a certain category by learning the relationship between input features and target classification. It is particularly suitable for discrete label classification tasks such as "positive/negative" in sentiment analysis.

3.2.5 Evaluation and Comparison

To measure the sentiment classification effect, the following indicators are used: Accuracy: Indicates the number of samples that the model predicts correctly, as a percentage of the total number of samples. In other words, "how many predictions are correct". TP is a true positive sample, TN is a true negative sample, FP and FN are false positive samples and false negative samples respectively. Accuracy is simple and intuitive, and is suitable for the situation where positive and negative samples are balanced. However, it may be misleading when the samples are unbalanced. For example, when the negative class accounts for the majority, even if the model predicts all negative classes, Accuracy may still be very high. Therefore, it is necessary to combine indicators such as Precision and Recall to comprehensively judge the performance of the model.

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision: It indicates the proportion of samples predicted as "positive" by the model that are actually "positive". TP is a true positive and FP is a false positive. The higher the Precision, the fewer false positives the model has and the more reliable the prediction. It is particularly suitable for scenarios with high false positive costs, such as spam identification. When the samples are unbalanced or the error type is more concerned, Precision is usually used together with Recall, and the overall performance of the model is comprehensively evaluated through F1-score to avoid one-sided judgment of a single indicator.

$$\text{Precision} = \frac{TP}{TP + FP}$$

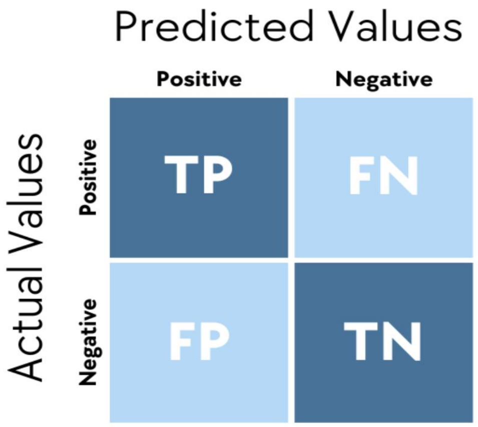
Recall: Represents the proportion of samples that are actually "positive" that are correctly identified by the model. TP is a true positive sample and FN is a false negative sample. The higher the Recall, the fewer positive samples are missed by the model, and the higher the coverage. It is very suitable for scenarios that are particularly sensitive to missed reports, such as disease detection and fraud identification. Recall and Precision usually need to be weighed and cannot be used alone. In order to comprehensively consider the accuracy and completeness of the model, F1–score is often combined to jointly evaluate the model performance.

$$\text{Recall} = \frac{TP}{TP + FN}$$

F1–score:It is the harmonic mean of precision and recall, and is used to weigh the overall performance of the two. The value of F1–score ranges from 0 to 1. The larger the value, the more balanced the model is between precision and coverage. When either Precision or Recall is very low, F1–score will also decrease. Therefore, it is suitable for tasks that require optimizing both indicators at the same time, such as text classification and sentiment analysis.

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Confusion Matrix:It is a table used to evaluate the effectiveness of a classification model, showing the correctness and errors of the model in predicting positive and negative classes. By comparing the predicted values with the true values, the model's correct (TP, TN) and incorrect (FP, FN) predictions on the positive and negative classes are shown.



		Predicted Values	
		Positive	Negative
Actual Values	Positive	TP	FN
	Negative	FP	TN

Log Loss Curve: In order to more comprehensively evaluate the model performance, in addition to static indicators such as accuracy, this article also records the Log Loss value that changes with the training round (epoch) during the model training process, and draws the Log Loss curve to observe the convergence trend and overfitting phenomenon. This indicator can reflect the performance of the model in terms of probability output and is more suitable for models with probability optimization objectives such as XGBoost.

3.3 Summary

This chapter introduces the overall methodological process of this study in detail. From the multilingual collection and preprocessing of review data to the construction and evaluation of sentiment models, all of them revolve around "The emotional differences of watching NEZHA 2 from a global perspective", ensuring that the analysis process is data-driven, the method is scientific, and the conclusions are credible.