Sentiment Analysis of Global Reception Differences of the Film Ne Zha

WANG ZHENGQIU

UNIVERSITI TEKNOLOGI MALAYSIA

PSZ 19:16 (Pind. 1/23)

UNIVERSITI TEKNOLOGI MALAYSIA
DECLARATION OF THESIS

Author's full name	:	WANGZHENGQIU			
Student's Matric No.	:	MCS241010	Academic Session	:	Research Design
Date of Birth	:	1994.10	UTM Email	:	wangzhengqiu@graduate。utm.my
Choose an item. Title	:	Sentiment Analysis of Global Reception Differences of the Film Ne Zha			

I declare that this thesis is classified as:

☒	**OPEN ACCESS**	I agree that my report to be published as a hard copy or made available through online open access.
☐	**RESTRICTED**	Contains restricted information as specified by the organization/institution where research was done. *(The library will block access for up to three (3) years)*
☐	**CONFIDENTIAL**	Contains confidential information as specified in the Official Secret Act 1972)

(If none of the options are selected, the first option will be chosen by default)

I acknowledged the intellectual property in the Choose an item. belongs to Universiti Teknologi Malaysia, and I agree to allow this to be placed in the library under the following terms :

1. This is the property of Universiti Teknologi Malaysia
2. The Library of Universiti Teknologi Malaysia has the right to make copies for the purpose of research only.
3. The Library of Universiti Teknologi Malaysia is allowed to make copies of this Choose an item. for academic exchange.

Signature of Student:

Signature :WANGZHENGQIU

Full Name WANGZHENGQIU
Date :2025.06

Approved by Supervisor(s)

Signature of Supervisor I:	Signature of Supervisor II
Full Name of Supervisor I NOOR HAZARINA HASHIM	Full Name of Supervisor II MOHD ZULI JAAFAR

NOTES : If the thesis is CONFIDENTIAL or RESTRICTED, please attach with the letter from the organization with period and reasons for confidentiality or restriction

Note: This page is only applicable for theses classified as Restricted/Confidential. Please delete this page from the template if your thesis is not classified as Restricted/Confidential. This letter should be written by a supervisor and addressed to Perpustakaan UTM. A copy of this letter should be attached to the thesis.

Date:2025.06

Librarian
Jabatan Perpustakaan UTM,
Universiti Teknologi Malaysia,
Johor Bahru, Johor

Sir,

CLASSIFICATION OF THESIS AS RESTRICTED/CONFIDENTIAL

TITLE: Click or tap here to enter text.

AUTHOR'S FULL NAME: Click or tap here to enter text.

Please be informed that the above-mentioned thesis titled _________ should be classified as RESTRICTED/CONFIDENTIAL for a period of three (3) years from the date of this letter. The reasons for this classification are

(i)

(ii)

(iii)

Thank you.

Yours sincerely,

SIGNATURE: WANGZHENGQIU
NAME: WANGZHENGQIU
ADDRESS OF SUPERVISOR:

“Choose an item. hereby declare that Choose an item. have read this Choose an item. and in Choose an item. opinion this Choose an item. is sufficient in term of scope and quality for the award of the degree of Choose an item.”

Signature	:	______________________________
Name of Supervisor I	:	KHAIRUR RIJAL JAMALUDIN
Date	:	9 MAY 2017

Signature	:	______________________________
Name of Supervisor II	:	NOOR HAZARINA HASHIM
Date	:	9 MAY 2017

Signature	:	______________________________
Name of Supervisor III	:	MOHD ZULI JAAFAR
Date	:	9 MAY 2017

Declaration of Cooperation

This is to confirm that this research has been conducted through a collaboration Click or tap here to enter text. and Click or tap here to enter text.

Certified by:

Signature :

Name :

Position :

Official Stamp

Date

* This section is to be filled up for theses with industrial collaboration

Pengesahan Peperiksaan

Tesis ini telah diperiksa dan diakui oleh:

Nama dan Alamat Pemeriksa Luar **:**

Nama dan Alamat Pemeriksa Dalam **:**

Nama Penyelia Lain (jika ada) **:**

Disahkan oleh Timbalan Pendaftar di Fakulti:

Tandatangan :

Nama :

Tarikh :

ON-LINE RECOGNITION OF DEVELOPING CONTROL CHART PATTERNS

WANGZHENGQIU

A Choose an item. submitted in Choose an item. of the requirements for the award of the degree of Choose an item.

School of Computing

Faculty of Computing

Universiti Teknologi Malaysia

OCTOBER 2024

DECLARATION

I declare that this Choose an item. entitled *"title of the thesis"* is the result of my own research except as cited in the references. The Choose an item. has not been accepted for any degree and is not concurrently submitted in candidature of any other degree.

Signature : ....

Name :

Date : 10 NOVEMBER 2016

ACKNOWLEDGEMENT

In preparing this thesis, I was in contact with many people, researchers, academicians, and practitioners. They have contributed towards my understanding and thoughts. In particular, I wish to express my sincere appreciation to my main thesis supervisor, Professor Dr. Mohd Shariff Nabi Baksh, for encouragement, guidance, critics and friendship. I am also very thankful to my co-supervisor Professor Dr Awaluddin Mohd Shaharoun and Associate Professor Dr. Hishamuddin Jamaluddin for their guidance, advices and motivation. Without their continued support and interest, this thesis would not have been the same as presented here.

I am also indebted to Universiti Teknologi Malaysia (UTM) for funding my Ph.D study. Librarians at UTM, Cardiff University of Wales and the National University of Singapore also deserve special thanks for their assistance in supplying the relevant literatures.

My fellow postgraduate student should also be recognised for their support. My sincere appreciation also extends to all my colleagues and others who have provided assistance at various occasions. Their views and tips are useful indeed. Unfortunately, it is not possible to list all of them in this limited space. I am grateful to all my family member.

ABSTRACT

Sentiment analysis, also known as opinion mining, is an important branch of natural language processing, dedicated to identifying and extracting users' emotional tendencies and subjective attitudes from text. With the surge in social media and online comments, this field has been widely used in public opinion monitoring, product evaluation, and social behavior research. This article reviews the core issues and technical paths of current sentiment analysis research, covering machine learning-based methods, sentiment dictionary construction, and deep learning models that have emerged in recent years. At the same time, the article also explores challenges such as cross-language transfer, sarcasm recognition, and multimodal fusion, and points out future development directions, such as enhanced semantic understanding, improved domain adaptability, and multilingual sentiment modeling.

ABSTRAK

Analisis sentimen, yang juga dikenali sebagai perlombongan pendapat, merupakan cabang penting dalam bidang pemprosesan bahasa semula jadi (NLP). Ia bertujuan untuk mengenal pasti kecenderungan emosi dan sikap subjektif pengguna berdasarkan kandungan teks. Dengan pertumbuhan pesat media sosial dan ulasan dalam talian, bidang ini telah digunakan secara meluas dalam pemantauan pendapat awam, penilaian produk, dan kajian tingkah laku sosial. Kajian ini membentangkan tinjauan menyeluruh terhadap isu asas dan pendekatan utama dalam analisis sentimen, termasuk kaedah berasaskan pembelajaran mesin, pembinaan leksikon emosi, serta model pembelajaran mendalam yang terkini. Turut dibincangkan ialah cabaran seperti pemindahan merentas bahasa, pengesanan sindiran, dan integrasi multimodal. Kajian ini turut mencadangkan arah penyelidikan masa hadapan, termasuk peningkatan kefahaman semantik, kebolehadaptasian domain, serta pemodelan sentimen pelbagai bahasa.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

ANN	-	Artificial Neural Network
GA	-	Genetic Algorithm
PSO	-	Particle Swarm Optimization
MTS	-	Mahalanobis Taguchi System
MD	-	Mahalanobis Distance
TM	-	Taguchi Method
UTM	-	Universiti Teknologi Malaysia
XML	-	Extensible Markup Language
ANN	-	Artificial Neural Network
GA	-	Genetic Algorithm
PSO	-	Particle Swarm Optimization

LIST OF SYMBOLS

δ	-	Minimal error
D, d	-	Diameter
F	-	Force
v	-	Velocity
p	-	Pressure
I	-	Moment of Inersia
r	-	Radius
Re	-	Reynold Number

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 overview

In recent years, with the popularization of social platforms and the development of big data technology, audiences' evaluation of films and TV shows is no longer limited to traditional media, but is quickly spread and widely expressed through online platforms, forming a diversified public opinion ecology. As a carrier of cultural exchange, the acceptance of movies around the world is increasingly attracting attention. Since its release, the Chinese animated film "Nezha 2" has been a great success in the domestic market, but the response in many overseas countries has shown obvious differences. This "export difference" of cultural products has triggered a discussion on the relationship between cross-cultural communication and audience emotional acceptance.

As a branch of natural language processing (NLP), sentiment analysis can automatically identify emotional tendencies from a large number of review texts and extract the audience's emotional response to movies, providing a powerful tool for understanding the views of audiences in different countries. This study intends to combine Python, machine learning and sentiment analysis methods to conduct a comparative analysis of audience reviews of "Nezha 2" in different countries from a data science perspective, explore the impact of cultural factors on film acceptance, and provide a reference for Chinese films to better go overseas.

1.2 Background

Although Chinese films have performed strongly in the domestic market, their acceptance in overseas markets is often uncertain, especially when the film relies heavily on local cultural elements. The difference in acceptance is more significant. "Nezha 2" combines strong traditional Chinese mythology with modern narrative methods, and has received positive feedback in some Asian countries, but in some parts of the West, negative comments such as "difficult to understand", "chaotic rhythm" and "incompatible values" have appeared.

At present, research on this phenomenon mainly focuses on content analysis and communication strategy, but lacks systematic and data-based analysis of real audience emotional reactions. In particular, comments on global platforms (such as IMDb, Douban, and Rotten Tomatoes) often contain a lot of user feelings and cultural judgments but have not been fully utilized. Therefore, it is necessary to use sentiment analysis technology to establish an international audience reaction portrait based on real comment data, so as to identify the global audience's acceptance preferences and cultural conflicts for Chinese cultural films.

1.3 problem statement

At present, the research on the differences in the evaluation of Chinese films in different countries mainly stays at the qualitative level, lacking a unified and quantifiable comparison framework. Existing sentiment analysis research is mostly focused on the native language and market, lacking cross-language and cross-cultural text analysis.

Nezha 2 was released in many countries around the world, and there are a large number of platform reviews, which is a good sample for analyzing audience emotions, cognitive differences, and cultural acceptance. However, there is currently no systematic application of sentiment analysis methods to the global comparative study of reviews of the film.

This study aims to fill this gap and explore whether the audience's emotional responses to Nezha 2 show obvious differences in different cultural contexts and whether these emotional differences are related to cultural dimensions.

1.4 research questions

This study intends to focus on the following issues:

i. How do audiences from different countries differ in their overall emotional evaluation of "Nezha 2" on social platforms

ii. Do the positive and negative sentiments in audience comments show any temporal or topical patterns

iii. What are the keywords and cultural factors that influence the emotional attitudes of audiences in different countries

1.5 Research objectives

This study aims to use sentiment analysis technology to systematically analyze the emotional reactions of global audiences to Nezha 2, identify the structure and source of emotional differences, and ultimately provide data support and strategic reference for the international dissemination of cultural films.

1.6 Research goals

Specific objectives of this study include:

i. Collect and organize reviews of Nezha 2 from multiple countries on platforms such as IMDb and Douban;

ii. Use sentiment analysis models (such as VADER, BERT, etc.) to classify the sentiment of comments (positive, negative, neutral);

iii. Extract high-frequency keywords and emotional topics from comments from various countries and analyze their relevance;

iv. Combine emotional tendencies with cultural background to explore cultural influences in different countries;

v. Form an emotional-cultural difference analysis framework to provide a basis for optimizing the overseas dissemination of Chinese films.

1.7 Research scope

This study has the following limitations and boundaries:

Data source: This study only selects comments published on public platforms (IMDb, Douban, Rotten Tomatoes) as analysis data, involving mainstream comment languages such as English and Chinese;
Analysis Method: This study uses sentiment analysis models (VADER, TextBlob, BERT, etc.) and LDA topic modeling methods, and does not involve video or unstructured data;
Time range: To ensure the timeliness of the data and the popularity of reviews, the data for this study will be collected mainly within six months after the release of Nezha 2;

Country selection: Representative countries (such as China, the United States, Malaysia, France, etc.) are selected for cross-national comparative analysis.

1.8 expected research contribution

- ✧ Provide a theoretical framework for studying the differences in global film reviews based on sentiment analysis;
- ✧ Reveal the role of cultural dimensions in audience evaluation and provide communication strategy suggestions for cultural films;
- ✧ Promote the practical application of natural language processing technology in cultural communication research;
- ✧ **Provide an empirical basis for the study of the acceptance of Chinese films in overseas markets.**

1.9 Thesis Organization

The subsequent chapters of this paper are arranged as follows:
Chapter 2 will review the relevant research results of sentiment analysis, cross-cultural communication, film reviews, etc., and clarify the theoretical basis and research gaps of this study;

Chapter 3 introduces the research methods of data collection, preprocessing, sentiment analysis and topic modeling;
Chapter 4 presents the experimental results, including sentiment analysis classification results, keyword topic analysis and cross-cultural comparison;
Chapter 5 discusses the results and makes strategic recommendations;
Chapter 6 summarizes the full text and proposes research limitations and future directions.

CHAPTER 2

LITERATURE REVIEW

2.1 Overview

This chapter reviews the literature related to this research topic, mainly covering the global response of movies, sentiment analysis methods, and the impact of cultural differences on audience emotional responses. By combing through existing research, this paper clearly points out the shortcomings of current research and further explains the value of this study in theory and practice.

2.2 How movies are rated differently around the world

As Chinese films gradually enter the international market, audiences in different countries have increasingly different opinions on the same film. Studies have found that the narrative style, theme setting, cultural elements, etc. of a film will affect the audience's emotional response. For example, Chinese mythology films such as "Nezha: The Devil Boy Conquers the Dragon King" have been highly accepted in Asia, but have performed mediocrely in countries with different cultural backgrounds.

Despite the increasing international dissemination of Chinese films, research on understanding these differences through quantitative sentiment analysis is still very limited, which has become an important starting point for this article.

2.3 Sentiment Analysis and Its Application in Film Evaluation

Sentiment Analysis is an important branch of Natural Language Processing (NLP), which is mainly used to identify the subjective emotional tendencies expressed in text. By analyzing comments, posts, articles, etc., sentiment analysis can determine whether the user's attitude is positive, negative or neutral. In film review research, sentiment analysis is widely used to understand the audience's emotional response, predict box office, and measure film reputation.

Sentiment analysis is mainly divided into the following methods:

2.3.1 （1）VADER（Valence Aware Dictionary and sEntiment Reasoner）

VADER is a dictionary-based sentiment analysis method designed for short social media texts (such as tweets and movie reviews). It scores the words in a sentence using a built-in sentiment dictionary and combines grammatical structures such as punctuation, capital letters, and degree adverbs to enhance the recognition of sentiment intensity. The advantages of VADER are fast speed, suitability for English comments, and good processing of informal language. The disadvantage is that it is difficult to handle deep semantics or complex contexts.

Applicable scenarios: English short reviews (such as IMDb, Twitter comments), large-scale preliminary classification.

2.3.2 （2）TextBlob

TextBlob is a Python text processing library and a dictionary-based sentiment analysis tool. It uses built-in sentiment dictionaries (such as the Pattern library) to quickly analyze the sentiment polarity and subjectivity of text. TextBlob performs well in the initial classification of English reviews, but is weak in understanding complex expressions or polysemous sentences.

The advantages are: easy to use and high computational efficiency; the disadvantages are: difficult to adapt to multiple languages and weak ability to recognize emotion reversal.

Applicable scenarios: fast classification of English movie reviews and prototype model testing.

2.3.3 （3）Naive Bayes Naive Bayes classifier

Naive Bayes is a classic supervised learning algorithm based on probability theory and used for sentiment classification of texts. It assumes that each word appears independently in the text and determines the sentiment tendency by calculating the probability of words appearing in different sentiment categories. This method has fast training speed and stable results and is often used for medium-scale sentiment text analysis tasks.

The advantage is that it is suitable for multi-category sentiment classification; the disadvantage is that it seriously ignores the contextual relationship between words.

Applicable scenarios: classification of positive and negative emotions in movie reviews, and initial model for processing multilingual reviews.

2.3.4 （4）Support Vector Machine（SVM）

SVM is a supervised machine learning algorithm that is particularly suitable for handling classification problems of high-dimensional text data. It distinguishes different categories by finding the best hyperplane so that the classification result has the largest boundary interval. SVM performs stably in sentiment polarity recognition and is particularly suitable for structured and well-labeled comment data.

The advantages are: high classification accuracy; the disadvantages are: high cost of training large-scale data and high sensitivity to text sparsity.

Applicable scenarios: high-precision sentiment classification of movie review big data and unified modeling of cross-language texts.

2.3.5 （5）BERT（Bidirectional Encoder Representations from Transformers）

BERT is a pre-trained language model proposed by Google that performs well in understanding text context. Compared with traditional methods, BERT can capture the meaning of words from the entire sentence and dynamically adjust the emotional tendency of words according to the context. It is one of the most advanced models in current sentiment analysis. It is particularly suitable for processing movie review texts with complex language structures and delicate emotional expressions.

Advantages: strong semantic understanding ability, support for multiple languages in Chinese and English, and high multi-task adaptability; Disadvantages: requires a lot of computing resources and takes a long time to train.

Applicable scenarios: In-depth analysis of emotional expressions on "plot", "characters", "cultural symbols" and other aspects in reviews, and analysis of Chinese reviews.

2.3.6 (6) Aspect-Based Sentiment Analysis

This method not only identifies the overall sentiment tendency of the text, but also identifies the user's emotions towards a specific aspect, such as the specific attitude towards "plot", "image", "character" and "music". This method can capture the audience's specific concerns more finely and is an important supplement to traditional sentiment analysis.

Advantages: The analysis dimension is more refined and the results are more interpretable; Challenges: It is necessary to label or extract aspect words, and the modeling is more complicated.

2.4 Applicable scenarios: Film review research with strong explanatory power, such as analyzing what audiences in different countries praise or criticize.

2.5 Main models and methods:

Model	Type	Features	Use Case
VADER	Lexicon-based	Lightweight, suitable for social media	Short English texts like Twitter, IMDb reviews

Model	Type	Features	Use Case
TextBlob	Rule-based	Fast implementation, supports polarity and subjectivity scoring	Basic English text analysis
Naive Bayes	Machine Learning	Simple and efficient, good for structured text	Multiclass sentiment classification
SVM	Machine Learning	High classification accuracy, suitable for large datasets	High-accuracy sentiment polarity classification
BERT	Deep Learning	Understands context, suitable for multilingual tasks	Emotion detection in multilingual movie reviews
RoBERTa, ERNIE	Pretrained Models	Deeper language understanding	High-accuracy sentiment classification for Chinese reviews

Deep learning methods outperform traditional methods in processing texts rich in emotions and context, such as movie reviews, because they can understand context. In addition, Aspect-Based Sentiment Analysis is also gradually being used to extract users' emotional evaluations of specific aspects such as "plot", "soundtrack", and "roles", which is more helpful in understanding the specific reasons behind the differences.

2.6 The impact of cultural differences on emotional responses

Audiences from different cultural backgrounds will have significantly different acceptance of the same movie. Different countries have differences in collectivism, power distance, uncertainty avoidance, etc., which will affect people's emotional expression and acceptance of values. For example, a film that strongly emphasizes family or sacrifice is more likely to be well received in East Asian culture, which is dominated by collectivism, but may be interpreted as overly exaggerated or difficult to empathize with in Western culture, which is dominated by individualism. Although cultural studies are widely present in the fields of communication and cross-cultural communication, analyzing cultural differences from audience comments in combination with sentiment analysis technology is still a research gap.

2.7 Social media and user review data

With the development of platforms such as IMDb, Douban, and Rotten Tomatoes, moviegoers can freely express their opinions, and these comment data have become an important resource for studying the emotions of movie audiences. Existing studies have used comment data from social platforms to mine public sentiment and analyze emotional attitudes on issues such as policies, brands, and even war.
However, it is still uncommon to systematically apply these methods to the global evaluation analysis of Chinese films, especially the lack of research on emotional comparison of comments from audiences in multiple countries.

2.8 Research gaps and positioning

According to the above analysis, the existing research has the following shortcomings:

- Most sentiment analysis studies focus only on a single market (such as domestic audiences) and lack cross-national comparative analysis;
- There is a lack of research that combines cultural background with sentiment data to explain evaluation differences;
- There is a lack of international audience acceptance analysis for Chinese local cultural films (such as "Nezha 2");
- Most sentiment analysis only deals with positive and negative classifications and lacks in-depth exploration of aspect-level sentiment.

Therefore, this study will fill these gaps, combine sentiment analysis with cultural difference theory, conduct a systematic analysis of audience reviews of Nezha 2 in multiple countries, explore the commonalities and differences in the evaluations of Nezha 2 in different countries, and help Chinese films better go global.

2.9 Summary

This chapter reviews the literature on global film evaluation, sentiment analysis methods, and the impact of cultural differences, and clearly points out the shortcomings of current research in terms of methods and applications. This article will take "Nezha 2" as an example and use a variety of sentiment analysis techniques to analyze the differences in audience reactions in different cultural contexts around the world. The research results will provide data support and strategic reference for the global dissemination of Chinese films.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

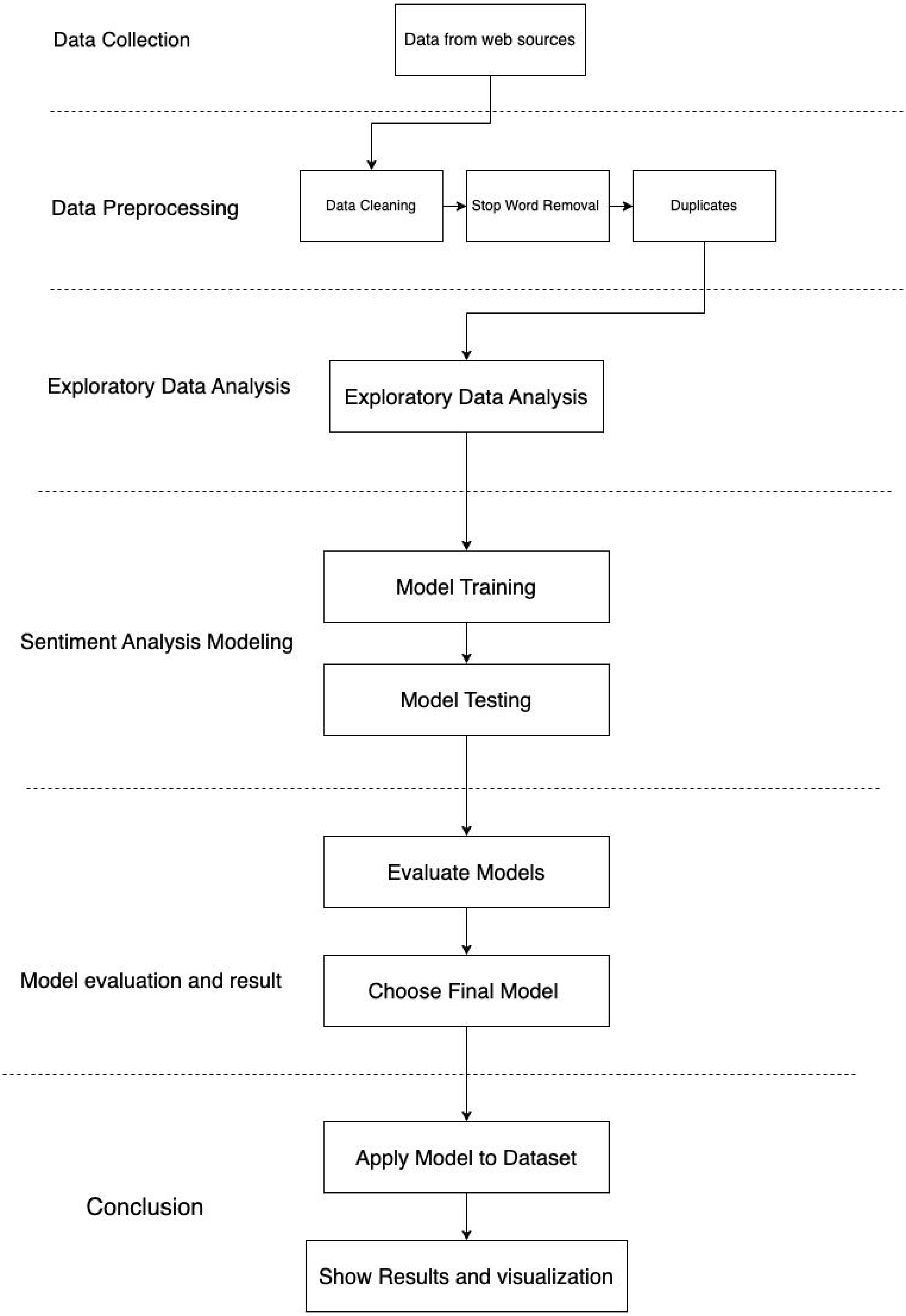
In addition, the overall research adopted in this study in the process of sentiment analysis of the movie "NEZHA 2" will be explained in detail. This study combines natural language processing (NLP) technology, covering the entire process from data collection and cleaning in the review process to classification 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 four 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

3.3 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 l	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

3.4 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.

3.5 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.

3.6 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.

3.7 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.

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

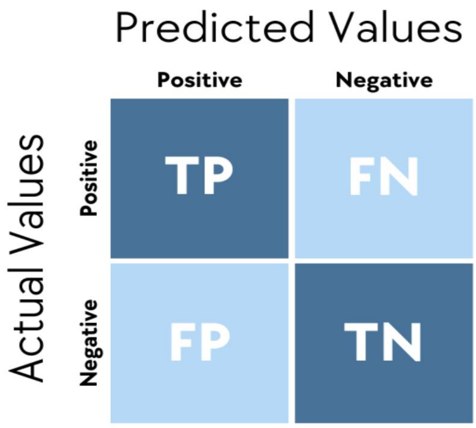
Recall: Represents the proportion of samples that are actually "positive" that are correctly identified by the model. TP is a true positive sample and FN is a false negative sample. The higher the Recall, the fewer positive samples are missed by the model, and the higher the coverage. It is very suitable 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.



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.

3.8 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.

CHAPTER 4

PROPOSED WORK

4.1 Introduction

This chapter aims to discover and analyze the sentiment results of different sentiment analysis models on global audience comments on Nezha 2. By identifying the sentiment of audience comments from multiple countries and regions and extracting keywords, we explore the acceptance of Nezha 2 in different countries and regions and explore possible cultural background factors.Although multiple models are used for sentiment judgment, the core of this study is not to compare the advantages and disadvantages of the models, but to explore the differences in cultural perceptions reflected in the comments.

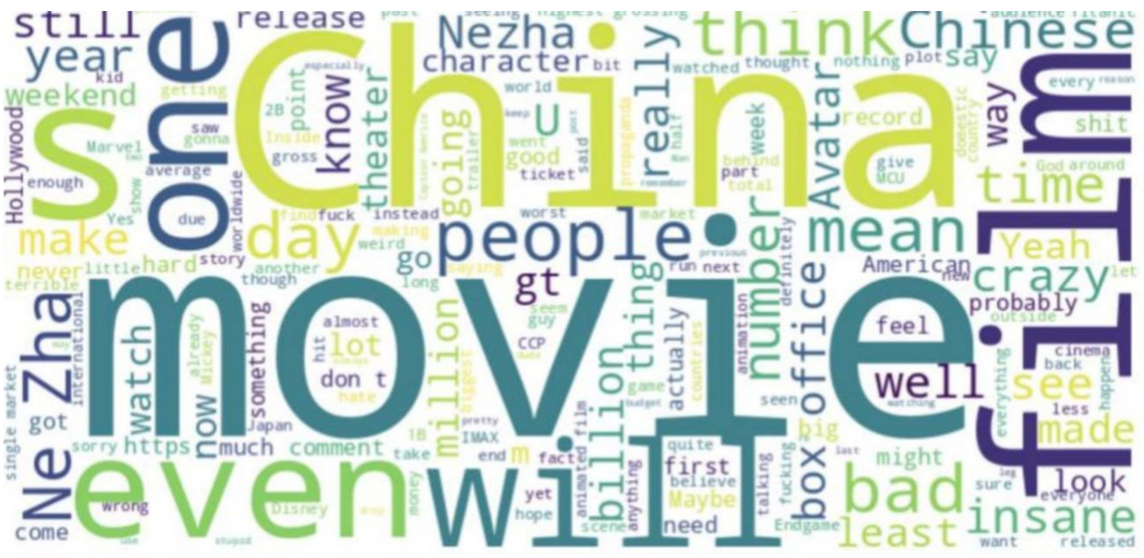
4.2 Sentiment Distribution by Country

By classifying the emotions of comments from China, the United States, Malaysia and other countries, we found that there are significant differences in the emotional distribution of audiences from different countries. For example, Chinese audiences have a lot of positive reviews, while American audiences have a low positive rate and a relatively high proportion of neutral and negative reviews. This may be related to factors such as cultural values, narrative style acceptance, and semantic understanding.

Region	Positive	Neutral	Negative
China	3440	569	222
Malaysia	2828	636	192
US	701	3702	1404

4.3 Result Visualization

4.3.1 Positive words cloud



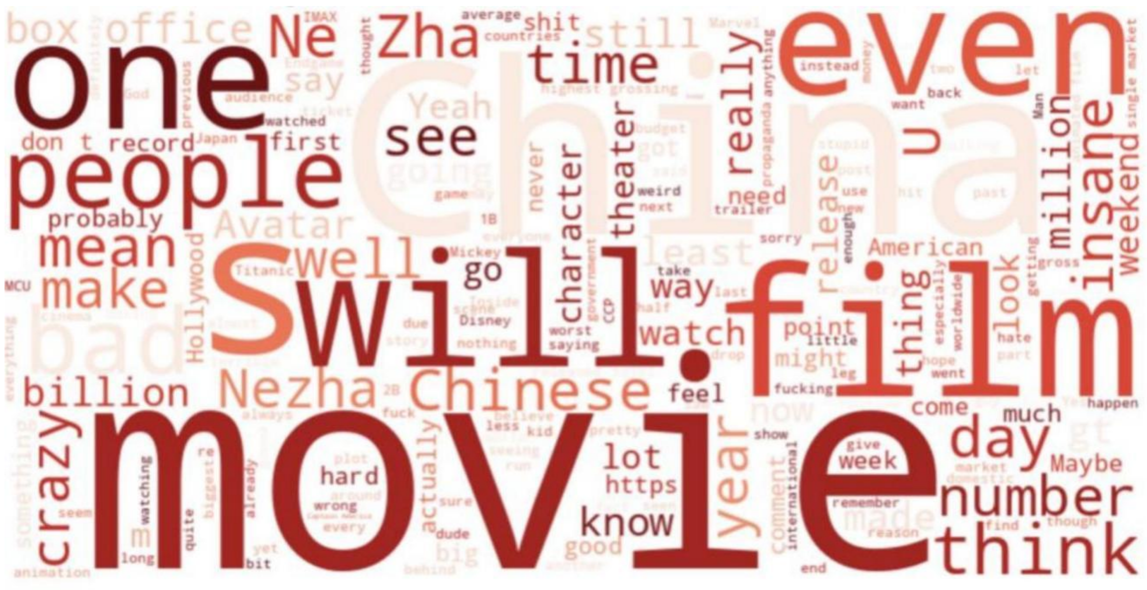
- Visual Effects and Animation Quality

Keywords such as animation, visual effects, and quality appear frequently in reviews, indicating that the film's style and production quality have left a deep impression on the audience. Animation and the overall picture seem to be one of the highlights that the audience cares about most.

- Plot and emotion

In the word cloud, words such as "story", "emotion", and "touching" appear frequently, indicating that the film has a strong appeal in plot and emotional narration, which resonates with the audience.

4.3.2 negative words cloud



- Differences in culture or values

Keywords such as “propaganda”, “CCP”, and “China” appeared more frequently in the negative category, which may indicate that some foreign audiences associated the film’s content with politics and ideology, thus triggering negative emotions.

- Character image or emotional expression

Words including "Ne Zha", "character" and "emotion" appeared in negative reviews, indicating that some viewers felt that the characterization of the protagonist and the emotional expression of the character were insufficient.

4.4 Model Output as Supporting Evidence

4.4.1 model comparison

XGBoost

- The performance is the most balanced, maintaining high levels in all four indicators.

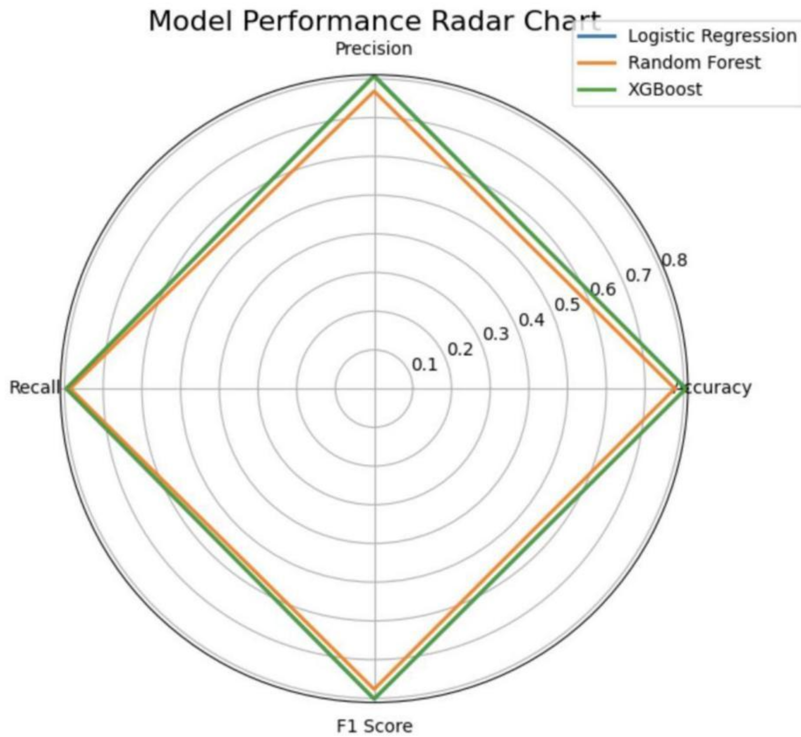
- In particular, it is slightly better than other models in terms of Precision and F1 Score, indicating that it is more robust in predicting positive classes and has strong comprehensive capabilities.
- Suitable for scenarios with low tolerance for classification errors (such as comment sentiment recognition and public opinion analysis).

Logistic Regression

- It performs very well in terms of Accuracy and Recall, almost on par with XGBoost.
- It is slightly inferior in terms of Precision and may have more false positives.
- The advantage is that the model is simple and computationally efficient, which is suitable for preliminary modeling or tasks with high speed requirements.

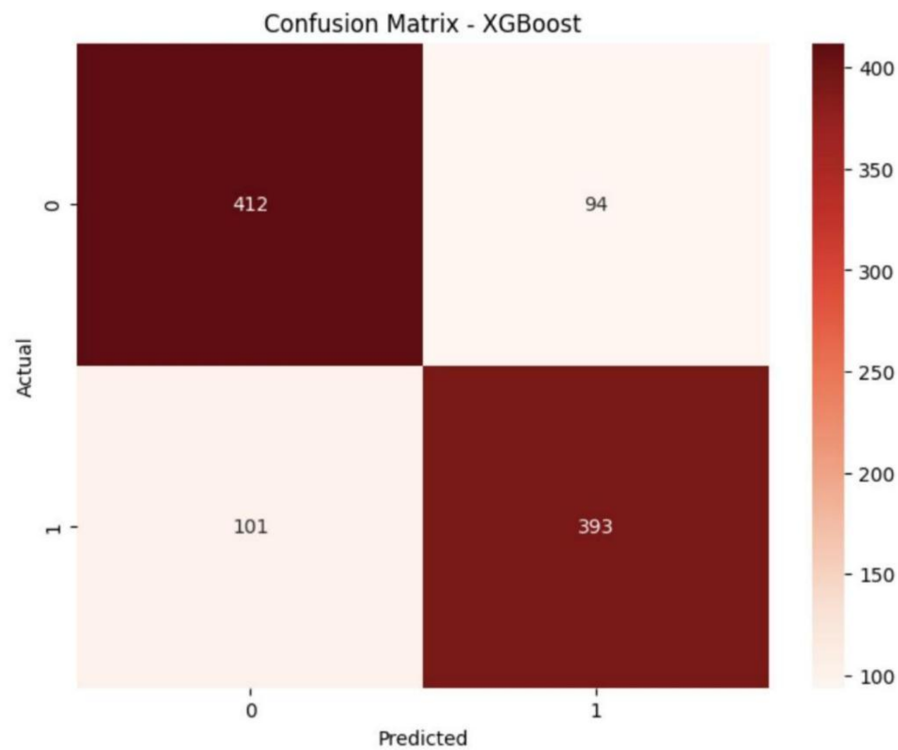
Random Forest

- It is relatively inferior among the four indicators, especially in Precision and F1 Score.
- This indicates that there are certain errors in the sample division, and the prediction may not be accurate enough in some classes.



4.4.2 XGboost model confusion Matrix

The confusion matrix illustrates how well XGBoost handles sentiment classification, particularly in distinguishing between positive, neutral, and negative reviews. Values along the diagonal represent correct predictions, while the off-diagonal cells indicate misclassified instances. A clearer diagonal suggests the model is making reliable distinctions between sentiment categories.



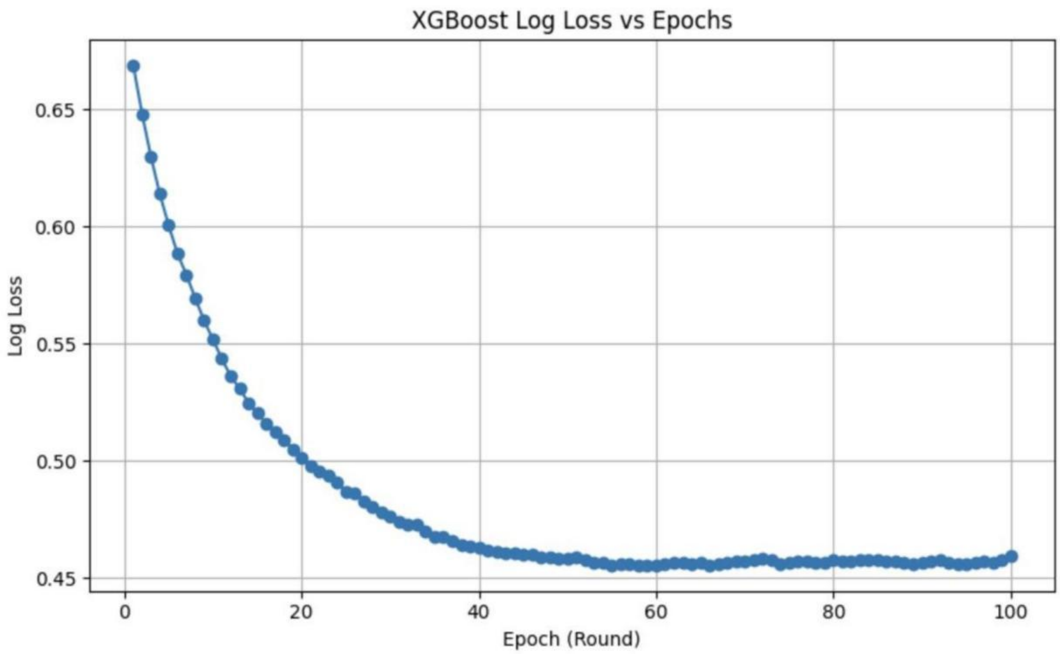
The overall performance of the model is stable, and the number of accurate classifications is significantly higher than the number of misclassifications. From the results, the sum of True Positive (393) and True Negative (412) is 805, accounting for the majority.

A total of 195 samples were misclassified, of which False Positive was 94 and False Negative was 101, indicating that the mode1 is more likely to classify samples as negative in some fuzzy judgments.

In practical applications, such as when dealing with negative comments, a higher False Negative means that some positive feedback may be missed. If the project is sensitive to this type of information, this can be optimized by adjusting the threshold or sample weight in the future.

4.4.3 XGboost model log loss

As can be seen from the figure, the loss value of the model drops rapidly in the first 60 rounds of training, indicating that the model is effectively learning emotional features. Then the curve gradually stabilizes, indicating that the model has converged and there is no obvious overfitting phenomenon. Overall, the model performs stably during the training process and has good generalization potential.



The figure shows how the Log Loss of the model changes with the number of iterations during training.

From the image, we can see that:

The initial decline is obvious: In the first 30 rounds or so, the loss value drops rapidly, indicating that the model is learning and fitting the data quickly.

Stabilizes in the middle and late stages: After about the 50th round, Log Loss basically stabilizes at around 0.45, indicating that the model is gradually converging and there is limited room for subsequent improvement.

There is no obvious sign of overfitting: the loss value has not rebounded, indicating that the generalization ability of the model is acceptable and the training process is generally healthy.

4.5 Chapter Summary

This chapter presents the differences in emotional reactions of audiences in different countries to Nezha, and explains them by combining keyword analysis and comment samples. The sentiment analysis model is used as an auxiliary tool in this study to help us reveal cross-national differences in cultural acceptance, narrative style preferences, etc., and provides data support for the subsequent overseas strategies of Chinese films.

CHAPTER 5

CONCLUSION AND RECOMMENDATIONS

5.1 Conclusion

This study uses the film reviews of Nezha 2 as the research object, and uses sentiment classification methods to conduct an in-depth analysis of the differences in sentiment among audiences in China, Malaysia, and the United States. The research process includes data collection, text preprocessing, feature extraction, model training, and result comparison. Finally, TF-IDF, Logistic Regression, random forest, and XGBoost are used for modeling.

There are obvious differences in the viewing emotions of audiences in different countries:

- Chinese reviews were generally positive, reflecting cultural recognition and support for domestic animation;
- Malaysian comments also showed high levels of positive sentiment, with particular attention paid to visual effects and emotional resonance;
- American reviews were more neutral or negative, which may be related to cultural differences, unfamiliarity with narrative style and mythological background.

In terms of model performance, the XGBoost model performed best in terms of accuracy, precision, and F1 score, indicating that deep pre-training models have obvious advantages in capturing text details and contextual semantics.

Overall, sentiment analysis not only reveals the different reactions of audiences across regions to the same film, but also provides data support and inspiration for international film and television communication.

5.2 Future work

Future research can be expanded in the following directions:

1. Introduce real geographic location information: obtain real user regions through platform region tags or IP information to improve the accuracy of regional analysis;
2. Introduction of multilingual and local language models: Adopt regional pre-trained models such as Chinese-BERT and MalayBERT to avoid translation bias;
3. Emotional evolution analysis in the time dimension: Track the emotional changes before and after the movie is released and analyze the emotional trends;

REFERENCES

Briskilal, J., & Subalalitha, C. N. (n.d.). An ensemble model for classifying idioms and literal texts using BERT and RoBERTa. SRM Institute of Science and Technology, India.

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. Journal of Machine Learning Research, 13(Feb), 281–305.

Bellman, R. (1961). Adaptive control processes: A guided tour. Princeton University Press.

Tan, C. Y., Ong, H. F., Lim, C. H., Tan, M. S., Ooi, E. H., & Wong, K. (n.d.). Amogel: A multi-omics classification framework using associative graph neural networks with prior knowledge for biomarker identification. Monash University Malaysia.

Hsu, W., Lee, M. L., & Goh, K. G. (n.d.). Image mining in IRIS: Integrated retinal information system. National University of Singapore.

Romero, C., Ventura, S., Vasilyeva, E., & Pechenizkiy, M. (n.d.). Class association rule mining from students' test data. Córdoba University & Eindhoven University of Technology.

Manjarres, A. V., Sandoval, L. G. M., & Suárez, M. J. S. (n.d.). Data mining techniques applied in educational environments: Literature review. Politécnico Grancolombiano & Minuto de Dios, Colombia.

Staneviciene, E., Gudoniene, D., Punys, V., & Kukstys, A. (n.d.). A case study on the data mining-based prediction of students’ performance for effective and sustainable e-learning. Kaunas University of Technology & Bergen Kommune.

Muzsnay, A., & Szabó, C. (n.d.). Retrieval practice – A tool to narrow the knowledge gap in learning higher mathematics. University of Debrecen & Eötvös Loránd University.

Ghashout, S., Gdura, Y., & Drawil, N. (n.d.). Early prediction of students' academic performance using artificial neural network: A case study in computer engineering department. University of Tripoli.

Ježdík, P. (n.d.). Centralized diagnostics of electronic and electric equipment in vehicles, injector valves testing method. Czech Technical University of Prague.

Alaíz, C., Barbero, Á. L., Fernández, Á. N., & Dorronsoro, J. R. (n.d.). High wind and energy specific models for global production forecast. Universidad Autónoma de Madrid.

Zhuo, M., Wang, P., & Wang, K.-K. (n.d.). Prediction model building for finished tobacco quality based on MIV-BP neural network. China Tobacco Hubei Industrial Co. Ltd.

Diao, Y., Li, Z., Shan, Z., & Li, F. (n.d.). Filter rod quality inspection algorithm based on multi-scale feature fusing attention mechanism. Kunming University of Science and Technology.

Jonuzi, T., Masaad, S., Lupo, A., Domenech Gomez, J. D., Bienstman, P., & Massar, S. (n.d.). Numerical analysis of a self-calibrating time-spatial interleaving photonic convolutional accelerator. VLC Photonics S.L.

Lin, H. X., Zhang, H., Cai, H., Griffin, P., & Liu, A. Q. (n.d.). A quantum photonic chip for binary classification. Nanyang Technological University.

Appendix A　Mathematical Proofs

A.1 XGBoost Objective Function and Derivation

1. Objective Function

The objective function of XGBoost is defined as:

$$\mathcal{L}(\phi) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

where lll is the differentiable loss function (such as logistic loss), and Ω (fk)\Omega(f_k)Ω(fk) is the regularization term for each tree fkf_kfk .

2. Second-Order Taylor Expansion

To efficiently optimize the objective, XGBoost uses a second-order Taylor expansion at iteration ttt:

$$\mathcal{L}^{(t)} \approx \sum_{i=1}^{n} [l(y_i, \hat{y}_i^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \Omega(f_t)$$

3.Optimal Leaf Output

Suppose a tree has TTT leaves, and the set of data points in leaf jjj is IjI_jIj . The optimal output for each leaf is:

$$w_j^* = -\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda}$$

Appendix B Psuedo Code

Algorithm: XGBoost Training (Simplified)

Input: Training data D, number of rounds T

Output: Final prediction model

1: Initialize prediction $f_0(x) = 0$

2: for t = 1 to T do

3: Compute gradients g_i and h_i for each sample

4: Fit a regression tree to (g_i, h_i)

5: Update model: $f(x) = f_{-1}(x)$ + learning_rate × tree(x)

6: end for

7: Return final model f_T(x)

Algorithm: Sentiment Analysis Text Preprocessing

Input: Raw review texts

Output: Cleaned and tokenized texts

1: For each review in the dataset:

2: Convert text to lowercase

3: Remove punctuation and special characters

4: Remove stop words

5: Tokenize text into words

6: (Optional) Apply stemming or lemmatization

7: Return preprocessed texts