

CHINESE MOVIE REVIEW BASED ON SENTIMENT
ANALYSIS AND MACHINE LEARNING

LI JINGYU

UNIVERSITI KEBANGSAAN MALAYSIA

CHINESE MOVIE REVIEW BASED ON SENTIMENT ANALYSIS AND
MACHINE LEARNING

LI JINGYU

PROJECT SUBMITTED IN PARTIAL FULFILMENT FOR THE DEGREE OF
MASTER OF DATA SCIENCE

FACULTY OF INFORMATION SCIENCE AND TECHNOLOGY
UNIVERSITI KEBANGSAAN MALAYSIA
BANGI

2025

ULASAN FILEM CINA BERASASKAN ANALISIS SENTIMENT DAN
PEMBELAJARAN MESEN

LI JINGYU

PROJEK YANG DIKEMUKAKAN UNTUK MEMENUHI SEBAHAGIAN
DARIPADA SYARAT UNTUK MEMPEROLEH IJAZAH SARJANA SAINS DATA

FAKULTI TEKNOLOGI DAN SAINS MAKLUMAT
UNIVERSITI KEBANGSAAN MALAYSIA
BANGI

2025

DECLARATION

I hereby declare that the work in this thesis is my own except for quotations and summaries which have been duly acknowledged.

10 February 2025

LI JINGYU
P132395

ACKNOWLEDGEMENT

This dissertation, in addition to the results of my intensive research about data science in my master's degree has been spiced up with the help and support from a group of very friendly people. Therefore, I write this paragraph to express my sincere gratitude and admiration to all of you!

First and foremost, I would like to express my sincere gratitude to my supervisor Prof. Dr. Azuraliza Abu Bakar, who has always listened to my ideas and given me direction on how to successfully implement my project. She took the time to meet with me even when she was busy and carefully suggested changes to each chapter. I am grateful for her openness to accept my ideas and give me the space to create my own and her generosity in providing guidance on feasible solutions when things get difficult. Moreover, I will thank all the professors who participated in the data science course. Their high-quality teaching has helped me master knowledge in the field of data science and has become the technical framework for building this project and providing theoretical support for feasibility assessment of project implementation.

Secondly, I would also like to thank Dr. Wandeeep Kaur, the coordinator of the Master of Data Science research report, for clearly presenting the important information about the project including the process, writing requirements, timelines, etc. in every briefing session and urging us to submit the milestones on time. This became an important basis for me to develop an efficient and high-quality project proposal and timeline. Furthermore, she was in charge of creating the Turnitin for each student's report, as well as providing the next steps once the results were in. Without their conscientious attitude, our project submission would not have gone as smoothly as it did, so I would like to thank her and the Qursiah team from the bottom of my heart.

Thirdly, I would like to thank the Faculty of Information Science and Technology of Universiti Kebangsaan Malaysia for its support and strong learning atmosphere for students to conduct academic research. The strong faculty team gives each student specialized training and has set up academic paper related courses to improve students' academic research ability.

Lastly, I would like to give a special thanks to my family for their constant support and love. You have been a strong motivation for me to move forward, providing unwavering encouragement and love when I was lost. It is because of your support that I have been able to research abroad and pursue my education without fear until now.

ABSTRAK

Analisis sentimen (SA) digunakan secara meluas dalam penyelidikan ulasan filem Cina. Ia bertujuan untuk menilai kepuasan penonton terhadap filem tersebut dengan menganalisis kecenderungan emosi dalam ulasan pengguna menggunakan teknologi pemprosesan bahasa semula jadi. Untuk tugas klasifikasi sentimen, kajian berkaitan telah menggunakan kaedah pembelajaran mesin tradisional seperti NB dan SVM dan model pembelajaran mendalam. Memandangkan terdapat hanya sedikit kajian tentang ketepatan penggunaan ML untuk mengklasifikasikan teg sentimen, penyelidikan ini akan menggunakan set data ulasan filem Cina, bertujuan untuk mencadangkan algoritma untuk menganalisis ulasan bahasa Cina dengan tepat, menggunakan TF-IDF untuk menukar teks kepada vektor ciri dan mengekstrak teg perasaan daripada data semakan, menggunakan model pembelajaran mesin berdasarkan analisis sentimen untuk menganalisis kecenderungan emosi penonton dalam data semakan filem Cina dan untuk menentukan kaedah klasifikasi sentimen terbaik dengan membandingkan ketepatan model klasifikasi ML. Penyelidikan ini mengaplikasi empat model klasifikasi ML yang berbeza: Mesin Vektor Sokongan (SVM), Naive Bayes (NB), Regresi Logistik dan Hutan Rawak (RF). Selepas eksperimen, ketepatan pengelasan empat model masing-masing ialah 0.77, 0.76, 0.78 dan 0.59. Antaranya, model Regresi Logistik mempunyai prestasi klasifikasi terbaik untuk data yang digunakan dalam penyelidikan ini. Cabaran yang dihadapi oleh penyelidikan termasuk semantik Cina yang kompleks, data jarang, dan generaliti model yang tidak mencukupi.

ABSTRACT

Sentiment analysis is widely used in Chinese movie review research. It aims to evaluate the audience's satisfaction with the movie by analyzing the emotional tendencies in user reviews using natural language processing technology. For sentiment classification tasks, related studies have used traditional ML methods such as NB and SVM and DL models. Since there are very few studies on the accuracy of using ML to classify sentiment tags, this research will use the Chinese movie review dataset, aims to propose an algorithm to accurately analyse Chinese language reviews, use TF-IDF to convert text into feature vectors and extract sentiment tag from review data, and use a ML model based on SA to analyse the emotional tendencies of audiences in Chinese movie review data and to determine the best sentiment classification method by comparing the accuracy of ML classification models. This research applied four different ML classification models: Support Vector Machine (SVM), Naive Bayes (NB), LR, and Random Forest (RF). After the experiment, the classification accuracy of the four models were 0.77, 0.76, 0.78 and 0.59 respectively. Among them, the LR model has the best classification performance for the data used in this research. The challenges faced by the research include complex Chinese semantics, sparse data, and insufficient model generality.

TABLE OF CONTENTS

DECLARATION		iii
ACKNOWLEDGEMENT		iv
ABSTRAK		vi
ABSTRACT		vii
TABLE OF CONTENTS		vii
LIST OF TABLES		x
LIST OF ILLUSTRATIONS		xi
LIST OF ABBREVIATIONS		xii
CHAPTER I	INTRODUCTION	
1.1	Introduction	1
	1.1.1 Related Research on Chinese Movie Reviews	1
	1.1.2 Sentiment Analysis Technology	2
	1.1.3 Related Research on SA Based on ML	2
1.2	Research Motivation	3
1.3	Research Background	4
1.4	Problem Statement	6
1.5	Research Questions	7
1.6	Research Objective	7
1.7	Scope of Research	8
1.8	Research Contribution	8
CHAPTER II	LITERATURE REVIEW	
2.1	Sentiment Analysis	11
	2.1.1 Overview of Sentiment Analysis	11
	2.1.2 Text Representation Tools in Sentiment Analysis (SA)	12
	2.1.3 Machine Learning Methods in Sentiment Analysis	13
	2.1.4 Deep Learning Methods and Sentiment Classification in Sentiment Analysis	13

2.1.5	Application Challenges and Development Directions of SA	14
2.2	Text Representation Methods	16
2.2.1	Bag of Words (BoW)	16
2.2.2	TF-IDF	19
2.3	ML Algorithm for Classification	22
2.3.1	Support Vector Machine (SVM)	25
2.3.2	Naive Bayes (NB)	27
2.3.3	Logistic Regression (LR)	30
2.3.4	Random Forest (RF)	32
2.3.5	Machine Learning to Analyze Sentiment Analysis	34
2.3.6	SA for ML Methodology	39
2.3.7	SA in Social Media Data	39
2.3.8	SA in Related Fields	40
2.4	Research Gap	42
2.5	Conclusion	43
CHAPTER III METHODOLOGY		
3.1	Introduction	45
3.2	Phase1: Data Collection and Development	46
3.2.1	Data Description	46
3.2.2	Data Pre-processing	49
3.3	Phase 2: Chinese Movie Reviews Sentiment Extraction	52
3.4	Phase 3: Sentiment Classification Using ML Classifiers	54
3.4.1	ML Algorithm	56
3.4.2	SVM Classification Model	57
3.4.3	NB Classification Model	57
3.4.4	LR Classification Model	58
3.4.5	RF Classification Model	58
3.5	Phase 4: Evaluation Methods	59
3.5.1	Accuracy	61
3.5.2	Recall	61
3.5.3	Precision	62
3.5.4	F1-Score	63
3.5.5	AUC - ROC Curve	63
3.6	Conclusion	64

CHAPTER IV	RESULTS AND EVALUATION	
4.1	Introduction	66
4.2	Results and Analysis of the SVM Model	66
4.3	Results and Analysis of the NB Model	69
4.4	Results and Analysis of the LR Model	71
4.5	Results and Analysis of the RF Model	74
4.6	Comparative Analysis of Four ML Classifiers	76
4.7	Comparison with Other Related Researches Using DL Methods	79
4.8	Conclusion	82
CHAPTER V	CONCLUSION	
5.1	Introduction	83
5.2	Result discussion	84
5.3	Limitation	85
5.4	Future Work	85
REFERENCE		88

LIST OF TABLES

Table No.		Page
Table 2.1	Summary of SA-related literature	15
Table 2.2	Summary of ML-related literature	41
Table 3.1	Movie Comments Chinese-English comparison table	47
Table 3.2	Movie Chinese-English comparison table	48
Table 3.3	Confusion Matrix	60
Table 4.1	SVM classifier classification report	67
Table 4.2	NB classifier classification report	69
Table 4.3	LR classifier classification report	71
Table 4.4	RF classifier classification report	74
Table 4.5	Classification report of each ML classifier for the positive class	77
Table 4.6	Classification report of each ML classifier for the negative class	77
Table 4.7	Accuracy comparison with related researches	81

LIST OF ILLUSTRATIONS

Figure No.		Page
Figure 2.1	Sigmoid function	31
Figure 3.1	Process of analyzing Chinese movie reviews	46
Figure 3.3	Comparison of the number reviews for different movies	48
Figure 3.4	Determination of emotional tendency in movie reviews	53
Figure 3.5	Classification of movie review sentiment labels	54
Figure 3.6	The proportion of emotions in different movies	55
Figure 3.7	Emotional distribution bar chart	56
Figure 3.8	Construction of SVM model	57
Figure 3.9	Construction of NB model	58
Figure 3.10	Construction of LR model	58
Figure 3.11	Construction of RF model	59
Figure 3.12	AUC-ROC curve	64
Figure 4.1	AUC-ROC curve of SVM model	68
Figure 4.2	AUC-ROC curve of NB model	70
Figure 4.3	AUC-ROC curve of LR model	73
Figure 4.4	AUC-ROC curve of RF model	76

LIST OF ABBREVIATIONS

CNN	Convolutional Neural Networks
DL	Deep Learning
DT	Decision Tree
FN	False negative
FP	False positive
LR	Logistic Regression
ML	Machine Learning
NB	Navie Bayes
NLP	Natural Language Processing
RF	Random Forest
RNN	Recurrent Neural Networks
SA	Sentiment Analysis
SVM	Support Vector Machine
TP	True positive
TN	True negative

CHAPTER I

INTRODUCTION

1.1 INTRODUCTION

1.1.1 Related Research on Chinese Movie Reviews

With the development of the Internet and the rise of social media, user-generated content (UGC) has exploded on much many online platforms, like social media, blogs, and comment areas, especially in the field of movie reviews. The expression of audience emotional attitudes has received unprecedented attention. Through the continuous in-depth research of predecessors, scholars have gradually realized the importance of text data. Yang (2013) pointed out in his research that there is a lot of valuable information hidden in text data. Therefore, more and more studies have begun to focus on how to use different algorithms to more deeply mine the information in text data, so as to discover more useful data. Chinese movie reviews are a typical example of this kind of emotional expression. Movie fans give feedback on movies through movie reviews and express their emotional attitudes and opinions on movies. Therefore, researching the emotional tendencies in Chinese movie reviews has become an increasingly important topic. S. Jayalakshmi et. al (2022) also pointed out that through the combination of sentiment analysis (SA) technology and machine learning (ML) algorithms, researchers can mine valuable information behind text data such as reviews and extract useful sentiment data. These comments can not only reflect the audience's emotional attitude towards the movie, but also reveal their opinions on the quality of the movie, the plot, the performance of the actors, etc.

Therefore, how to deeply explore the emotional information in these movie reviews are of great significance to the development of the movie industry.

1.1.2 Sentiment Analysis Technology

SA technology has a history of more than 30 years. As the Internet and social media develop more and more swiftly, it is gradually become an important tool to analyse the text data. The core goal of SA is to identify the tendency of sentiment (such as Positive, Negative and Neutral) from the text and further reveal the emotional information hidden behind the text. As users continue to post comments and opinions about various events or products online, especially on social platforms, SA technology can effectively capture these emotional attitudes and analyse them. According to the research of Li et. al (2022), SA technology has broad application prospects in business, society, politics and other fields. For example, in the business field, SA can help companies better understand consumers' emotional attitudes towards products, and then adjust marketing strategies and product designs; in the social field, Yang (2013) pointed out that through the SA of user comments online, it is possible to better know the public's emotional trends and provide support for crisis management and public relations activities; in the political field, SA can also help governments and political parties better grasp voters' emotions and attitudes, and optimize elections and policy making. The research of Lv et. al (2021) pointed out that although SA technology has broad application prospects in various fields, due to the complexity and variability of emotions, as well as cultural and language differences, SA still faces many challenges when processing text data.

1.1.3 Related Research on SA Based on ML

The combination of SA and ML methods is become a hot research area in academia. According to Abbasi et. al (2021) said that, the application of ML algorithms in SA has been increasingly valued by scholars, especially when traditional SA methods face

difficulties. The introduction of ML methods has significantly improved the effect of SA. The research of Sidi et. al (2023) pointed out that researchers have combined related methods such as ML and SA. For example, researchers such as Wankhade et. al (2022), have extracted text features by using methods like BoW and TF-IDF, and combine traditional ML algorithms such as Navie Bayes (NB), Support Vector Machine (SVM) and Linear Regression (LR) aim to classify sentiment tags. These methods can improve the accuracy and robustness in SA, and are particularly effective in sentiment classification and label recognition. In addition, the research of Fayyaz et. al (2020) pointed out that with the introduction of Deep Learning (DL) technology, the effect of SA has been further improved. However, due to the complexity and variability of human emotions and the large differences in language and culture, SA still faces challenges in processing text data. Hassan (2017) pointed out that DL methods, especially word vector technology (such as Word2Vec and GloVe), can effectively capture semantic information in text and improve the accuracy of SA. At the same time, Li et. al (2023) pointed out that deep neural networks (such as CNN, RNN, LSTM) have shown excellent performance in processing long texts and complex semantic tasks. In terms of pre-trained language models, the application of models such as Transformer, BERT and GPT further improves the effect of sentiment classification because these models can better understand the deep semantic information in the context. Liu (2023) pointed out that combining SA with ML technology can help researchers extract sentiment information from text more accurately, thereby providing more accurate sentiment data support for the personalized recommendation system of movie review data.

1.2 RESEARCH MOTIVATION

As the continuous advancement of SA related technologies and the increasing number of comments posted by users on certain things or products on the Internet, the information hidden behind the comments generated by users has become more valuable for research. Mining the potential information contained in user comment

data on online network platforms has also become a hot research direction in various fields. Therefore, this research conducts SA based on ML methods on the movie review data posted by users on the Chinese movie online platform, aiming to explore users' sentiment polarity and viewing preferences for Chinese-language movies.

Fulzele et. al (2023) pointed out that the growth of the Chinese movie market and its global influence continue to expand, and there is a substantial demand for analyzing and predicting movie trends. With the development of big data analysis and AI technology, in-depth mining of movie review data can provide decision support for all links in the movie industry chain. Internet users and consumers also continue to post their comments on events or products on online platforms such as social media to express their emotional attitudes, which has led to an explosive growth trend in UGC on the Internet. For researchers, Furtado (2020) pointed out that the rational use of these text data and the use of SA-related technologies and ML algorithms can effectively discover the useful information behind these comments and other text data, which has broad application prospects in various fields of society, including movie recommendations.

1.3 RESEARCH BACKGROUND

The user-generated data on the Internet is also growing exponentially. Therefore, it is easier and more willing for all sectors of society to use text information such as comments and opinions published online. By applying different computer technologies, such as ML algorithms and SA related technologies to these text information data, different purposes can be achieved. For example, Alrumaih et. al (2020) pointed out that, for relevant government departments, by collecting and monitoring public opinion information on different online platforms and performing SA on public opinion data, the public's emotional attitude trends can be grasped in a timely manner to ensure a healthy social speech atmosphere and achieve timely detection and avoidance of risks. For producers, Haw (2022) pointed out by collecting

and analysing the comments data of consumers on a certain product published on different online shopping platforms and SA, consumers' views and satisfaction with the product can be accurately grasped, which helps producers provide products or services that can better satisfy consumers and increase commercial profits. Therefore, more and more scholars have begun to conduct SA on tweet data such as comments in the network and online platforms to identify customers' emotional tendencies and understand their real needs.

DL technology such as Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) have better processing effects on more complex and nonlinear problems. Therefore, some scholars like Tembhrne and Diwan (2020), have begun to combine neural network models with SA-related technologies. He et. al (2023) pointed out that, this development has greatly improved the processing effect of SA-related technologies on large-scale data and high-dimensional data and has achieved better results than traditional single algorithms when used for click-through rate prediction and user portrait construction. Similarly, Liu (2023) said applying SA-related technologies to the analysis of audiences' review data for a specific movie to recommend movies that will greatly facilitate users' lives. This makes it possible to analyse users' viewing preferences and generate personalized movie recommendations for users, greatly saving users' time and improving efficiency. However, Das and Singh (2023) pointed out, there are challenges in analyzing only based on SA-related algorithms, such as the sparseness and imbalance of sentiment data and the complexity of text processing. This research focuses on applying ML-related algorithms, like SVM, NB, LR and RF, to analyze audience sentiment tendencies in movie review data, and conducts a detailed research of Chinese movie review data based on related SA technologies.

1.4 PROBLEM STATEMENT

The analysis of audience movie reviews, particularly in the Chinese context, faces multifaceted challenges that hinder comprehensive sentiment understanding and predictive accuracy.

Firstly, Zhang et. al (2022) pointed out that the fragmented nature of movie review data on different platforms such as Douban, Weibo, and TikTok brings significant complexity to data processing due to heterogeneous formats and unstructured text. While recent studies have explored cross-platform data aggregation tools, however, Li et. al, (2021) presented that scalable solutions for harmonizing multilingual and multimodal user-generated content remain underdeveloped, especially for colloquial expressions and regional dialects prevalent in Chinese reviews.

Secondly, as highlighted by Wang et. al (2023) in their analysis of Chinese social media data, the linguistic complexity of Chinese movie reviews, such as irony, ambiguity, and cultural differences, poses a huge obstacle to natural language processing (NLP). Although pretrained language models like BERT have advanced sentiment analysis (SA), their performance deteriorates when handling informal or context-dependent phrases.

Furthermore, traditional sentiment analysis frameworks struggle to capture the dynamic interplay of socio-psychological factors influencing audience emotions, such as personal preferences and cultural biases. Chen et. al (2023) demonstrated that static models fail to account for evolving emotional attitudes during movie viewing, limiting their utility in real-time feedback prediction. Additionally, cross-cultural variations in movie interpretation and the prevalence of ironic language further complicate SA tasks. Recent work by Liu et. al (2022) revealed that even state-of-the-art multimodal models exhibit poor generalizability in detecting sarcasm across demographic groups.

This highlights the urgent need to find algorithms suitable for Chinese movie review data to improve the accuracy of audience sentiment analysis and prediction.

In this research, the most important thing is to compare the accuracy of different ML classification algorithms for the classification of user emotional labels in movie review data, aiming to select the ML algorithm that best fits the data used in this research, that is, to select the ML classifier that produces the best classification results for emotional labels. The optimal classifier can better grasp the emotional information of the audience in the movie review data. It is worth mentioning that the best ML classifier selected in this research may not be applicable to other data sets.

1.5 RESEARCH QUESTIONS

This research uses ML algorithms to perform SA on Chinese movie review data, aiming to more accurately identify users' sentiment polarity towards related movies, better identify users' viewing preferences, and maximize the effect of movie recommendations in the future. The researches in this research are:

1. RQ1: How improve the algorithm to accurately analyse Chinese language reviews?
2. RQ2: How to extract accurate sentiment from the reviews data?
3. RQ3: How to model the sentiment labelled data to perform accurate classification?

1.6 RESEARCH OBJECTIVE

The research objectives are as follows.

1. RO1: To propose an algorithm to accurately analyse Chinese language reviews.
2. RO2: To use TF-IDF to convert text into feature vectors and extract sentiment tag

from review data.

3. RO3: To select the best ML classifier for the Chinese movie review data.

1.7 SCOPE OF RESEARCH

This research collected the movie review data of Chinese movies from Douban.com users on the open data website Kaggle. These data are limited to user reviews in China. These movie review data cover twenty eight different kinds of Chinese-language movies. Each movie contains the user's rating data and evaluation data for the movie. The necessary data pre-processing and feature engineering are performed on the Chinese-language movie review data using Python. This research focuses on SA based on different ML algorithms for the movie review data of Chinese-language movies posted by users, and according to the user's rating of the movie divides the user's emotional tendency into 'Positive' and 'Negative'. The specific ML algorithms involved are Support Vector Machine (SVM), Logistic Regression (LR), Navie Bayes (NB) and Random Forest (RF). The method that best suits the data used in this research is determined by comparing the classification accuracy of these four ML classifiers for sentiment labels. Since the method used in this research is only based on the movie review data of Chinese-language movies on Douban.com, the data source is relatively single, so the results of the research may not be applicable to movie review data from other online platforms, and the definition of sentiment labels may also have a certain impact on the performance of the model.

1.8 RESEARCH CONTRIBUTION

The contribution of this research is as follows: First, this research collected Chinese movie review data from different online platforms such as Douban.com on the open data platform Kaggle, integrated and summarized them, and obtained the audience's ratings and review data for Chinese movies of different themes and types. Secondly, this research proposed a combined algorithm based on ML algorithm to perform SA

on Chinese movie review data to judge the audience's emotional attitude towards movies. The SA method combined with ML classifier has a high accuracy rate for movie recommendations, can better analyse the audience's emotional attitude towards a certain theme of movies, and provides new ideas for the research of Chinese movie review data. Finally, for movie producers, they can adjust movie production and market positioning strategies according to the results of algorithm analysis to improve the market competitiveness of movies; for distributors and exhibitors, they can formulate more reasonable scheduling and publicity plans based on the prediction model; for audiences, they can obtain personalized movie recommendations through recommendation algorithms to improve the viewing experience. For researchers and policymakers, providing quantitative data support for the development of China's movie industry will help promote the healthy development and international competitiveness of China's movie industry.

The progress for this research is as follows; Chapter 1 mainly introduces the relevant research background and methods of Chinese movie reviews, and points out the problems, research questions and objectives, research scope and the contribution value of this research. Chapter 2 mainly studies and analyses the different methods used by predecessors in the analysis of Chinese movie reviews, introduces the principles of ML algorithms such as SVM, NB, and LR, as well as the related research and applications in the field of Chinese movie review analysis and SA, and points out the significance of SA based on ML algorithms for researching Chinese movie reviews. Chapter 3 mainly introduces the relevant steps and methods such as data preprocessing, model construction and evaluation performed in this research when analysing movie review data. Chapter 4 describes and analyses the results of the methodology in the previous chapter, compares the accuracy of classifying sentiment tags using different ML classifiers, selects the best ML classification model based on the data of this research, and reviews and summarizes the entire project. Chapter 5 reviews and summarizes the entire research, provides solutions to the research

problems raised, and points out the shortcomings of this research and suggestions for future improvements.

LIBRARY FTSM

CHAPTER II

LITERATURE REVIEW

2.1 SENTIMENT ANALYSIS

2.1.1 Overview of Sentiment Analysis

Sentiment Analysis (SA) is widely used in Natural Language Processing (NLP) technology, aiming to identify, judge and extract positive, negative or neutral emotional information expressed in text. Kumar et. al (2024) also pointed out that the classification and prediction accuracy of ML algorithms in big data processing and complex problem solving has been significantly improved. Therefore, SA has become an important tool for extracting emotional information from large amounts of data and has been widely used in different research fields. In-depth research on SA can help companies and research institutions obtain more valuable information when understanding and analysing public sentiment and promote the development of related fields. SA not only plays an important role in movie recommendations, but is also widely used in other fields such as business and medical care. Zheng (2022) collected disease feedback data from Chinese patients, they used the jieba tool for text data preprocessing and used SA technology to generate personalized treatment recommendations for patients. Their research pointed out that the problem of missing sentiment vocabulary in SA can be solved by more precise data preprocessing technology or model optimization technology.

2.1.2 Text Representation Tools in Sentiment Analysis (SA)

Text representation methods such as BoW, n-grams, TF-IDF, etc. in SA play an important role in improving sentiment classification performance. Qasem and Sajid (2022) combined BoW and n-grams with TF-IDF for fake news detection and proved that these text representation methods can significantly improve the effect of text classification models. However, the research also pointed out that the problems of high-dimensional feature space and insufficient semantic depth require more complex models to solve. Similarly, Hadi and Utami (2024) explored the method of combining BoW, TF-IDF and n-grams with the KNN algorithm to classify hate speech. They believed that this method is helpful in handling text data classification tasks, but it also faces the problems of insufficient semantic depth and sparse features.

García et. al (2021) studied the application of n-grams in text classification and explored the construction efficiency of n-grams and the impact of feature selection techniques on text classification performance. The research emphasized the advantages of n-grams in capturing contextual information and pointed out that in the application of large-scale or complex data sets, dimensionality and sparsity are challenges that need to be paid attention to. Mounika et. al (2021) studied the combination of CNN and n-grams for SA of book reviews. The research showed that combining n-grams with CNN can enhance the ability of SA, but it also faces problems such as increased computational cost and overfitting. Habberrihat et. al (2023) conducted SA, using TF-IDF and n-grams for feature extraction and combining ML techniques to analyse text data from central Libya. The research showed that combining TF-IDF and n-grams with ML models can significantly improve the accuracy of sentiment classification, but it also faces problems such as high-dimensional feature space and insufficient semantic understanding.

2.1.3 Machine Learning Methods in Sentiment Analysis

Research on SA often uses ML methods, such as NB, DT, SVM, etc. to classify sentiment information. Pavitha et. al (2022) crawled movie review data from the IMDB website and used these ML methods to analyse the sentiment in the reviews. After comparison, they found that the SVM classifier had the highest accuracy (98.63%), significantly higher than NB (97.33%). However, this research has certain language barriers and cannot analyse reviews other than English, and there may be errors in the classification of sarcastic or mocking reviews. Another common SA method is Latent Dirichlet. allocation (LDA).

Zhang et. al (2022) analysed the Chinese movie recommendation algorithm, using the LDA model to extract topics from movie reviews on Douban.com and combined it with SA technology to generate a movie recommendation list. Although this method can effectively combine text mining and recommendation systems, the expressiveness of the recommendation model still has room for improvement due to the neglect of sentiment polarity and feature intersection. In addition, VADER is a rule-based SA tool that is particularly suitable in processing informal texts such as social media and news comments. Kumar et. al (2022) used the VADER method to analyse movie-related reviews from Twitter and Weibo. The results showed that VADER has a good effect on short text sentiment recognition and can help generate personalized recommendations. However, the limitation of their research is that it is based on a static database, and it is still possible to explore the application of dynamic paradigms in the future.

2.1.4 Deep Learning Methods and Sentiment Classification in Sentiment Analysis

Deep Learning (DL) technology has shown strong advantages when processing complex text data. For example, Qaisar (2022) used LSTM for SA, and achieved a classification accuracy as 89.9% on the IMDB movie review dataset. Their research

provides another effective solution for the sentiment polarity classification task. Kim et. al (2022) used a dictionary-based SA method to identify adjectives in movie reviews, to improve the accuracy of the review classification. They quantified movie reviews into sentiment scores and used SVM, RF and NNnet. algorithms for sentiment classification. Their research showed that adding adverbs to the sentiment dictionary can refine the sentiment expression and further improve the accuracy of the recommendation system. Ullah et. al (2022) published a research on SA of movie reviews using deep neural networks (DNNs) in the journal Complexity. The research highlights the advantages of using advanced DNN models to classify sentiment in user-generated content, specifically focusing on their application to movie reviews. The research adopted a DNN-based approach for sentiment classification, using word embeddings and DL architectures to analyse and classify movie reviews, and integrated techniques such as word2vec and pre-trained embeddings to enhance feature representation, ensuring that the model can capture the context and semantic nuances of the reviews. Compared with traditional ML methods (such as NB, SVM), the DNN model showed excellent performance and higher accuracy and robustness when processing large datasets. While the research showed significant progress in the accuracy and scalability of deep neural networks in enhancing SA of movie reviews, the research also highlighted challenges in terms of computational requirements, data dependency, and interpretability.

2.1.5 Application Challenges and Development Directions of SA

Although SA methods have made a lot of progress, they still face certain challenges in practical applications. For example, Chatterjee et. al (2022) collected movie review data from the IMDB website and used ML and NLP methods for SA. However, since the dataset is almost entirely positive reviews, there is an imbalance problem in sentiment classification, which can be solved by adding more negative reviews. In addition, the use of DL models has great potential in processing SA. Dellal-Hedjazi et. al (2022) combined LSTM with SA and used LDA for topic modelling to analyse the

sentiment information in Amazon review texts. Although this method improves the accuracy of SA, it also faces problems such as high dimensionality and insufficient semantic understanding.

Table 2.1 Summary of SA-related literature

Author	Dataset	Methods	Topic	Gap
Pavitha etc. (2022)	IMDB web movie review data	NB, DT, SVM and SA	Compare the accuracy of ML classifiers in labeling the sentiment information in movie reviews as positive and negative.	Language barriers and cannot correctly classify sarcastic or ironic comments.
Zhang (2022)	Douban web movie review data	Latent Dirichlet. allocation (LDA) and SA	Using LDA to mine potential topics in text and combining it with SA to improve classification accuracy.	The sentiment polarity of a specific topic is not taken as a feature, the intersection of features is not considered, and the model lacks expressiveness.
Kumar (2022)	Twitter Tweets Data	VADER	Use VADER to identify the sentiment polarity of short text tweet data and control user sentiment information.	Exploration in a dynamic paradigm has possibilities.
Chatterjee (2022)	Self-scraped review data from IMDB web	RF, XGBoost, SVC and SA	Comparing the accuracy of ML classifiers in classifying the sentiment in the data.	Since the dataset is almost completely positive, there is an imbalance in to be continued...

continuation...

				sentiment.
Kim (2022)	NAVER Movies movie review data	SA Based on Dictionary, SVM, RF and NNet. algorithm	Use the dictionary-based SA to identify adjectives in movie reviews and compare the accuracy of the results of the classification methods used.	The construction of the sentiment dictionary lacks adverbs and is insufficient in the recognition of detailed sentiment knowledge.

2.2 TEXT REPRESENTATION METHODS

2.2.1 Bag of Words (BoW)

As Hasan.T (2021) mentioned that BoW is a fundamental text representation method that converts unstructured text into structured numerical vectors while ignoring grammatical order. Its core idea is to treat a document as an unordered collection of words (a "bag"), focusing only on whether and how frequently words appear rather than their sequence or contextual relationships. The specific Implementation Process as follows.

1. Vocabulary Construction. Extract all unique words from the corpus to form a vocabulary. For example, given the sentence 1 "The cat sat on the mat" and sentence 2 "The dog ate my homework", the vocabulary becomes [the, cat, sat, on, mat, dog, ate, my, homework].
2. Vector Encoding. For each document, create a vector with the same length as the vocabulary. Each element in the vector records the frequency of the

corresponding vocabulary word in the document. The process is, sentence 1: [2, 1, 1, 1, 1, 0, 0, 0, 0], Sentence 2: [1, 0, 0, 0, 0, 1, 1, 1, 1].

The implementation principle of the BoW model is to represent each document as a vector and the length of the vector is equal to the size of the vocabulary in the document. The text vectorization formula in equation 2.1.

$$v_d = [w_1, w_2, \dots, w_n] \quad \dots(2.1)$$

Among them, w_i is the weight of the i -th word in the document can be TF, TF-IDF or binary features (if word i appears, it is 1, if word i does not appear, it is 0), indicating the size of the constructed vocabulary containing the number of unique words. Generally, the larger the data set, the more unique words it contains, and the larger the vocabulary. Text vectorization is to convert text into vectors to provide reasonable and algorithm-recognizable input data for subsequent ML and similarity calculations.

When using the BoW model to represent text, the relationship between documents can be measured by calculating the similarity between vectors. Usually, cosine similarity and Euclidean distance are used to calculate the similarity value of vectors. The formula involved is as follows in equation 2.2 and 2.3.

1. Cosine similarity

$$\cos(d_1, d_2) = \frac{v_{d_1} \cdot v_{d_2}}{\|v_{d_1}\| \|v_{d_2}\|} \quad \dots(2.2)$$

Among them, $v_{d_1} \cdot v_{d_2}$ is the dot product between two documents represented by vectors, $\|v_d\|$ is the magnitude of the vector.

2. Euclidean distance

$$\text{dist}(d_1, d_2) = \sqrt{\sum_{i=1}^n (w_{i,d_1} - w_{i,d_2})^2} \quad \dots\dots(2.3)$$

Among them, w_{i,d_1} , w_{i,d_2} are the weights of the i -th word represented by the vector in documents d_1 and d_2 respectively.

By using cosine similarity to calculate the similarity between vectors, we can measure the topic similarity in documents. For example, we can compare whether two articles of different lengths discuss the same topic. That is, although the length of different documents is quite different, as long as the proportion of vocabulary in the two documents is the same, the two articles will be considered similar. By using Euclidean distance to calculate the similarity between vectors, we can measure the feature differences between texts. For example, we can detect whether the absolute difference between different documents, such as the number of words or vocabulary, is significantly different. In other words, if the number of a particular word in one document is much higher than that in another document, this absolute difference will be reflected by the Euclidean distance. Usually, in SA, when we process documents, we pay more attention to the content of the document rather than its length. If a document with a modifier added expresses the same topic as the original document, the modulus of the document vector after the modifier is added will increase. If only Euclidean distance is used to calculate similarity, longer documents may have a larger distance due to the existence of longer document vectors, thereby masking the similarity between the actual content of the documents. Therefore, using cosine similarity for calculation can ignore the difference in document length and only focus on the similarity between contents, effectively avoiding the problem of only considering document length and ignoring the consistency of document content.

N-grams are commonly used in natural language processing to capture local contextual relationships at the character and word level. As pointed out by Qasem A.E. (2022), extending N-grams based on the BoW model can better capture the local word

order information of the text. Its implementation principle is to divide the text into a sequence of consecutive n characters or words. The formula is in equation 2.4:

$$n - \text{grams}(d) = \{(w_i, w_{i+1}, \dots, w_{i+n-1}) \mid i \in [1, |d| - n + 1]\} \quad \dots(2.4)$$

Among them, w_i is the i -th word in the document, n is the number of consecutive characters or words extracted from the text each time, that is, the size of the window. In other words, if $n=1$, it means extracting a single word or character each time, which means that the context is not considered; if $n=2$, it means extracting two consecutive words or characters from the text each time to capture the dependency relationship between two adjacent words; when $n=3$, it means extracting three consecutive words or characters from the text each time, which means that a longer context relationship can be captured. When $n>3$, a longer sequence of words or characters can be extracted, which is suitable for tasks such as specific complex grammatical analysis that need to capture richer context relationships. By using N -grams sequences based on BoW, the BoW model can better identify word order information in documents.

2.2.2 TF-IDF

It is an important text feature representation method, which is an indicator to measure the importance of the word in the text. It combines the frequency of a word or character in a specific document with the prevalence of the word in the entire corpus used to help us extract the characters or words that best represent its characteristics in the current document. Among them, 'TfidfVectorizer' in sklearn library is used to calculate the weight of the word. TF is term frequency, which is used to measure the frequency of a character or word in a document, and IDF is inverse document frequency, which is used to measure the rarity of a character or word in the entire corpus. Generally speaking, a character or word with a higher TF-IDF value means

that it is an important keyword in the document. The formulas involved are shown in equation 2.5 and equation 2.6.

$$TF(t, d) = \frac{\text{The number of times word } t \text{ appears in document } d}{\text{The total number of words in document } d} = \frac{f(t,d)}{\sum_{t \in d} f(t,d)} \quad \dots\dots(2.5)$$

$$IDF(t) = \log \frac{N}{1+df(t)} \quad \dots\dots(2.6)$$

Where N is the total number of documents, t is the word, and $df(t)$ is the number of documents containing word t . The TF-IDF formula obtained by combining the TF and IDF formulas is in equation 2.7:

$$TF - IDF(t, d) = TF(t, d) \times IDF(t) \quad \dots\dots(2.7)$$

As pointed out by Liu et. al (2022), TF-IDF has excellent performance in text classification and information retrieval tasks for small-scale corpora and in resource-constrained environments. However, since it cannot capture the contextual relationship of words, for example, the sentiment polarity of "not bad" and "bad" is completely different, but it cannot distinguish them. Therefore, for some words with the same or opposite meanings, these rare sentiment words can only be highlighted when using this method for calculation, but they cannot be distinguished. In other words, in sentiment classification tasks, although TF-IDF cannot distinguish sentiment words, the feature weights it assigns to characters or words can help the model more accurately distinguish positive and negative sentiments in documents, and using TF-IDF values as input features for ML classification models (such as SVM, LR, etc.) can improve the performance of classification models.

It should be noted that Sundaram et. al (2021) pointed out that TF-IDF has other limitations: It is sensitive to the length of the text, because a word or character may have a higher frequency in a long document and is more dominant in the weight calculation, so additional normalization is required during the calculation. In addition,

since the value of IDF is based on a static corpus containing a fixed data set built when performing NLP tasks, since the content of the corpus is no longer updated or changed after it is built, the feature weights of the model need to be recalculated when the corpus changes.

In general, as Liu et. al (2022) pointed out, when using TF-IDF for SA tasks such as text classification and information retrieval tasks, it is necessary to combine n-grams or other complex language models to enhance its expressiveness and capture phrase-level contextual dependencies, which helps to better overcome the problem of semantic and contextual information not being recognized. By using BoW and TF-IDF, text data can be converted into numerical vectors that can be recognized by ML models.

Kauffmann et. al (2020) pointed out, as a natural language processing technology, SA has certain commercial value, social value, academic value and cultural research value. It provides individuals, enterprises and social organizations with powerful tools to gain insight into consumer behaviour patterns, optimize service quality and predict economic trends by analysing and mining the emotional information contained in text, voice or images. Therefore, it has a wide range of applications in many areas of life, such as monitoring public opinion sentiment in social media such as Twitter and Facebook, and analysing customer feedback data, which is used to process customer evaluation text information on products, helping companies better understand customer preferences and needs to improve products. For market research, Sanchez.N et. al (2020) pointed out, SA technical analysis can help consumers provide feedback and guidance on products and services, thereby establishing marketing strategies. So, whether it is commercial application or social research, SA can help researchers understand the important role of emotions in decision-making and analysis more deeply and thoroughly.

2.3 ML ALGORITHM FOR CLASSIFICATION

Fahle et. al (2020) pointed out that ML is the main research field of AI at this stage and the training effect of ML-related algorithms on large amounts of data has been clearly improved in classification accuracy. Therefore, using ML to deal with large amounts of data and solve complex problems has been widely favoured by scholars in various research fields.

Sarma et. al (2021) pointed out ML algorithm is a method of processing data by using different algorithms and statistical models to automatically identify the patterns and information contained in the data from a large amount of training data and make corresponding decisions such as classification or prediction tasks. The basic principle of ML-related algorithms is to use the used or fitted data to learn the characteristics of the data. In this learning process, the algorithm can automatically obtain data feature information from the data by adjusting the corresponding hyperparameters of the model, so as to achieve more accurate prediction or classification tasks for new data. As Athar et. al (2021) pointed that commonly used supervised learning algorithms are SVM, RF and so on. And as Dellal and Alimazighhi (2022) pointed common unsupervised learning algorithms include Principal Component Analysis (PCA), K-means Clustering and so on.

Through the above introduction to the principle of ML fitting data, it can be seen that ML can achieve the established goal by automatically extracting data features, and extracting, classifying or predicting the useful information or emotional tendencies contained in the data from a large amount of data (including text data). Therefore, this research analyses and explains the research conducted by previous researchers using ML algorithms and the research on SA of text data based on ML as follows.

Juluru et. al (2021) explored the application of the BoW technique in the field of NLP with a particular focus on its use for radiologists. Their research provided an accessible primer on BoW, highlighting how this simple and effective method can be leveraged for analyzing clinical texts and medical reports. They highlighted the utility of BoW as an entry-level technique for text analysis in radiology, demonstrating its role in simplifying the handling of clinical data for ML applications. While BoW provides a solid foundation for text processing tasks, its limitations such as the loss of context, potential for high dimensionality, and insufficient semantic analysis must be addressed when employing it in real-world, complex medical data environments. Advanced NLP techniques, such as word embeddings or transformer models, may be required to complement BoW for better contextual understanding and more accurate predictions in medical applications.

Handlan (2020) aimed to leverage advanced ML techniques to better understand the impact of language used in official FOMC (Federal Open Market Committee) statements on financial markets and economic predictions. Their research demonstrated how ML can be applied to text analysis for understanding FOMC statements and predicting monetary policy impacts. The use of ML techniques provided quantitative insights that can support economic forecasting and policy analysis. However, the research also highlighted several limitations, such as challenges in capturing context, the dependence on data quality, and interpretability of model results. Hassan et. al (2022) conducted a comprehensive research to analyse the use of various ML algorithms for text classification, exploring their efficacy, applications, and comparative performance. Their research is significant for understanding how different ML models can be leveraged for tasks like SA, spam detection, and topic categorization, which are essential for processing and extracting meaningful information from large text datasets. The research provided valuable insights into the comparative performance of various ML algorithms for text classification, showing how certain models are better suited for specific types of data

or tasks and highlighted the role of feature extraction methods like BoW and TF-IDF in preparing data for analysis and discussed how these preprocessing steps influence model effectiveness. However, limitations such as scalability, feature dependence, interpretability, data quality, and handling of imbalanced datasets underscore the need for careful model selection and tuning for specific applications.

Luo (2021) presented an in-depth analysis of various ML techniques for efficient English text classification, exploring their implementation, performance, and suitability for different types of text data. Their research effectively demonstrated the potential of various ML techniques for English text classification, emphasizing the importance of model choice and feature extraction methods and the findings provide a solid foundation for understanding which ML models perform best under certain conditions, but the limitations in scalability, feature engineering, model interpretability, and data preprocessing underscore areas for further research. Poornima and Priya (2020) conducted a comparative analysis of SA using ML techniques, focusing on the efficacy of sentence embeddings for enhancing text classification tasks. Their research explored how different ML algorithms interact with various sentence embedding methods to decide which approaches shows better in SA tasks. However, their limitations, including potential data constraints, scalability issues, and the lack of DL exploration.

Okoye et. al (2022) explored the use of text mining and ML for analysing SETs in an educational context. By developing a contextual model, the research aimed to provide deeper insights into student feedback, allowing for improvements in teaching strategies. However, the research's limitations, such as dataset specificity, interpretability issues, and potential scalability challenges. Harrison and Sidey-Gibbons (2021) provide a practical introduction to the use of ML in the field of medicine, specifically focusing on the applications of NLP. Their research provided a guide on the application of ML and NLP in medical research, outlining the potential

benefits and challenges involved. While ML can enhance data processing, improve diagnosis, and support evidence-based medical decisions, its implementation is constrained by data quality issues, interpretability concerns, regulatory requirements, and technical barriers.

2.3.1 Support Vector Machine (SVM)

As Pisner and Schnyer (2020) pointed out that SVM is a supervised learning method, often used for classification, regression and anomaly detection tasks, the core idea is to find a can maximize the interval between two classes of data points of optimal hyperplane to best tag data classification, the method for high and with complex boundary classification problem is more applicable. The meaning of classification interval is the distance from the sample point closest to the hyperplane to the hyperplane, and the expression is like equation 2.8.

$$\text{Margin} = \frac{2}{\|w\|} \quad \text{.....(2.8)}$$

Mustafa and Mohsin (2021) pointed out, the task of SVM is to achieve the purpose of classification by maximizing the classification interval. Maximizing the classification interval is equivalent to minimizing the distance between the two types of sample points on the boundary of the classification interval, so that $\frac{1}{2}\|w\|^2$ can achieve the minimum value and satisfy the classification constraints, among them, the determination of support vectors is particularly important because it is vital in finding the optimal hyperplane. Arefinia et. al (2022) pointed out that, for linearly separable data, SVM will find a hyperplane that maximizes the classification interval, that is, the minimum distance between the two types of sample points used in the data and the hyperplane. The hyperplane can be expressed as equation 2.9.

$$w \cdot x + b = 0 \quad \text{.....(2.9)}$$

Where, w is the normal vector, x is the input feature, b is the bias. Since it is difficult to find a straight line to classify data in low-dimensional space, it is necessary to introduce kernel techniques. As Hasan et. al (2020) pointed out, kernel techniques are defined as mapping data to high-dimensional space by introducing kernel functions when data is indivisible in low-dimensional space, making it linearly separable. Kernel function refers to the function used to calculate the inner product of two sample points in high-dimensional space, expressed as equation 2.10.

$$K(x_i, x_j) = \phi(x_i) \cdot \phi(x_j) \quad \dots\dots(2.10)$$

Among them, $\phi(\cdot)$ is the mapping function from the original space to the high-dimensional space. By introducing the kernel function, SVM and other algorithms can solve nonlinear problems in high-dimensional space. As Roman et. al (2021) pointed out that common kernel functions include linear kernel, polynomial kernel, Gaussian kernel (RBF kernel), Sigmoid kernel, Laplace kernel, and histogram intersection kernel. Negi et. al (2024) emphasized that when choosing kernel functions, it is necessary to combine different data sets and usage situations.

For the problem of linear non-separable data, Hristopulos (2024) pointed out, the data can be mapped to a higher dimensional space by introducing Kernel tricks to make it linearly separable in a higher dimensional space. Among them, w is the weight, x is the input, b is the bias, and the sign function represents the classification result. The SVM decision function involved as follows in equation 2.11.

$$f(x) = \text{sign}(w \cdot x + b) \quad \dots\dots(2.11)$$

In general, after the text is vectorized using text feature extraction methods such as TF-IDF, the classification mechanism of the SVM model can classify the emotional attitude in the document, thereby helping to better analyse the emotional