澳 門 科 技 大 學

MACAU UNIVERSITY OF SCIENCE AND TECHNOLOGY

學 士 學 位 作 業 報 告

《基於自然語言處理和回歸算法的餐廳在線評論情感分析和預測》

學生編號	學生姓名
1909853V-B011-0151	郭子涵
17098533-B011-0187	潘尚德
1909853F-B011-0241	孫昊洋
1909853L-B011-0035	趙雨桐

學　　院：商學院

課程名稱：工商管理學士學位課程

指導老師：姜慧敏

遞交日期：2023 年 4 月 5 日

摘要

隨著在線美食平臺的流行，越來越多的用戶傾向於在這類平臺上尋找餐廳，並留下關於用餐體驗的評論。對於餐飲從業者來說，這些評論起到了監督的作用；對於消費者來說，他人的評價能夠起到輔助決策的作用。這促進了餐廳服務的良性發展。因此，餐飲從業者和數據分析人員對用戶評價的量化愈加迫切，如何做好評價指標量化是餐飲業當前所面臨的問題。

本文旨在探討餐飲服務業中使用自然語言處理技術與回歸算法對用戶評論進行情感分析的可行性。本文提出了兩個模型，分別為基於 TF-IDF 特徵提取與邏輯回歸的機器學習模型，以及基於 BERT 與 Softmax 分類層的深度學習模型，並在使用美團 ASAP 數據集進行訓練之後，對兩個模型的分類性能進行評估，最終得到餐廳評論語境下表現較佳的情感分類模型。為了保證輸出結果的客觀性，我們將同時考慮評論文本的整體情感與 5 個方面類別情感，從多方面體現評論文本所包含的情感信息，從而更好地幫助消費者與商家根據評論做出決策。

在考慮訓練集樣本分佈情況對模型性能的影響後，本文使用原始數據與過採樣處理後的數據，分別基於不同的建模方法，訓練多個不同的模型並對其性能進行分析。實驗結果顯示，基於 BERT 與 Softmax 分類的深度學習模型在不使用過採樣方法的情況下表現最佳，優於同條件下的機器學習模型。最終，我們選出表現最佳的一組模型，實現了餐廳評論整體情感與方面類別情感的預測與輸出，有助於消費者與商家對評論情感信息的挖掘與綜合考慮。

本文的方法和結果能夠為餐飲企業提供更好的用戶評論管理和服務改進的依據，把握用戶喜好和消費需求，並以此來對餐廳服務與菜品進行改善；同時也能夠幫助消費者更加有效快速地瞭解商家的優缺點，找到心儀的餐廳就餐。

Abstract

With the popularity of online food platforms, more users tend to look for restaurants on such platforms and leave comments about their dining experience. For the restaurants, the reviews are able to supervise effectively; For consumers, others' reviews are helpful in decision-making. These phenomena promote the healthy development of restaurant service. Therefore, catering practitioners and data analysts are more necessary to quantify users' reviews, and the quantization of evaluation indicators is the problem we face in the catering industry.

This paper aims to explore the feasibility of using natural language processing technology and regression algorithm for restaurants reviews' sentiment analysis. In this paper, two models are proposed: the machine learning model based on TF-IDF feature extraction and logistic regression, and the deep learning model based on BERT and Softmax classification layer. After training with Meituan's ASAP dataset, the classification performance of the two models will be evaluated, and we will choose the better sentiment classification model between them. To ensure the objectivity of the output results, we will simultaneously consider the sentiment of the review text and the five detailed categories about restaurants' services and food mentioned in the review, then reflect the sentiment information contained in the review text from various aspects. It will be helpful to consumers and restaurants making decisions based on the review.

After considering the influence of sample distribution on the model performance, this paper uses the original data and the oversampled data to train several different models with different algorithms, then analyze their performance. The experimental results show that the deep learning model based on BERT and Softmax classification has the best performance without oversampling processing, which is superior to the machine learning model under the same conditions. Finally, we selected a group of models with the best performance to realize the prediction and output of the overall sentiment and the category sentiment of restaurant reviews, which is helpful for consumers and restaurants to mine and comprehensively consider the sentiment information of reviews.

The methods and results of this paper can provide a better basis for catering enterprises to manage user reviews and improve service, grasp user preferences and consumption needs, and improve restaurant services and foods. It can also help consumers understand the advantages and disadvantages of restaurants, then find their favorite ones.

目錄

1. 前言

1.1 研究的背景與意義

餐飲行業是一個非常龐大、多元化的產業。從傳統的餐飲文化到現代的快餐、西餐、日韓料理、泰國菜等各種風味，涵蓋了不同消費者群體的口味偏好。中國的餐飲市場目前呈現出多元化的發展趨勢，包括餐飲連鎖化、品牌化、特色化及高端化。餐飲行業競爭非常激烈。

與此同時，消費者對餐飲質量、服務、環境等方面的要求也不斷提高，消費者更加注重健康、營養、安全等方面的因素。因此，現代餐飲企業也開始注重品牌建設、形象塑造、服務體驗等方面，追求客戶滿意度和忠誠度。

根據《2022 中國餐飲行業生態白皮書》[1]，美團表示，在過去兩年中，中國餐飲企業持續加速連鎖化和線上化，市場的連鎖化率從 13% 上升至 18%，增長了 5 個百分點。疫情的結束帶來許多不確定性，消費習慣和場景也隨之發生變化，只依靠傳統堂食消費的經營模式現已過時。為了提升效率和增強抗風險能力，餐飲企業必須進行數字化轉型和升級。據中國連鎖經營協會的調查表明，超過 68% 的頭部連鎖餐飲企業已認識到數字化轉型的重要性。美團這種公開透明的餐飲評價平臺將成為企業構建“以消費者為中心”這種模式的重要信息來源，同時也將為消費者提供重要的參考。

隨著信息技術和社會的不斷進步，人們的生活越來越智能化，特別是在餐飲領域，出現了眾多如美團、大眾點評一類的應用程序。人們可以在線上瀏覽和篩選餐館，也可以進行評論。很快，這一領域就實現了用戶數量的爆炸式增長，產生了數量龐大的消費評論數據集。對於這些數據集的情感分析任務，也就是如何從中找出用戶的真實感受，成為了一個重要問題。如果我們能準確地進行情感分析，剔除錯誤數據，這將有助於推動餐飲行業的進步和創新，真正的理解用戶的深層需求，為消費者和商家找到一條共贏的道路。

情感是人類智能的一個重要表現，它對我們的日常生活、學術交流、人際交往和決策都有著深遠的影響。研究表明，情感影響著人們的信息交流行為。當今社會，隨著互聯網的蓬勃發展，人們的網絡行為和信息傳遞中往往都包含著用戶想表達的情感信息，並互相傳播。因此，為了更好的分析用戶需求，挖掘評論文本與其背後的情感之間的關係成了一個重要的研究課題。

[1] https://ent.cnr.cn/canyin/zhiku/20220812/t20220812_525961896.shtml

分析大量非結構化文本數據是一項具有挑戰性的任務。自然語言處理 (NLP) 是人工智能的一個子領域，是分析和理解文本數據的有力工具。情感分析是一種自然語言處理的任務 (赵妍妍, 秦兵, & 刘挺, 2010)，需要從給定的文本評論中提取實體，分析實體中所包含的情感，屬文本分類的一種。它的目的是通過分析文本評論的褒貶性來確定文章的情感傾向。情感分析在餐飲、服務、娛樂等各個領域都有極為廣泛的應用。隨著在線餐飲服務的普及和用戶數量的增加，用戶評論成為了獲取消費體驗信息的重要來源，尤其是更具參考價值的用餐後評價。然而，隨著用戶數量的不斷增長，手動收集和分析這些評論已經變得低效、難以完成。因此，自動化分析用戶評論已經成為了不可或缺的一種解決方案。消費者可以從歷史評論中瞭解其他消費者的用餐感受，以避免不必要的麻煩；商家則可以利用這些信息瞭解自身服務的優劣和缺陷所在，以便及時調整並提供更好的服務。

參考此前的研究，我們發現對於評論情感分析，常見的做法是對評論文本與用戶給出的評分進行情感分析與建模。但經過觀察，我們發現用戶給出的評分存在一定的主觀因素：在多條評論內容相似的樣本中，客戶給出的評分存在較大差異。此外，根據日常生活中的經驗，我們推測，現實中可能會出現評論內容為負向情感，但評分因為某些原因給了高分的情況。片面地根據評價分數對評論文本的情感進行分析，並不利於用戶與平臺對餐廳的服務進行分析、判斷與監督，也不利於店家根據用戶反饋，對餐廳服務進行改進。

1.2 研究的方法和目的

現有的研究已經在餐飲行業進行了評論情感分析，並輸出了相應的模型。但另一方面，這些研究的目標是為了實現更好地對現有數據進行擬合，而非便於普通商家和用戶進行應用。另一方面，由於在實驗中使用了大量的數據及較高算力的設備，普通用戶可能無法使用現有的設備進行模擬或複現。因此，本文提出了一種能夠高效完成多方面類別評論情感分析與預測的方法，且兼顧精確度與計算可行性，在使用兩個維度的判斷指標進行模型評估的同時，減少用戶使用時的學習成本。

本文基於自然語言處理技術和回歸算法，使用美團 ASAP 數據對餐廳評論的情感進行建模與預測。我們將結合評論文本本身的情感傾向與一系列體現餐廳服務質量、具有正負向情感傾向的方面類別標簽作為用戶情感的體現，提供一個相對客觀的評論情感分析模型。我們從用戶的評論文本出發，分別對文本的情感和其提及的方面類別進行分析建模，最終得到一系列用於預測評論情感和“餐廳評價標簽情感”的模型。

本研究提出的模型能夠為餐飲行業管理者和消費者提供更加科學、客

觀、準確，同時又具有低門檻、低成本特點的分析和預測手段。對於餐廳管理者來說，能夠高效分析顧客反饋與需求，更好地瞭解和掌握市場動態，及時進行改進和優化，提高服務質量與用戶滿意度；對於消費者來說，可以通過餐廳評論的情感分析更加準確地瞭解產品和服務的質量和特點，從而做出更明智的消費決策。

2. 相關研究成果

2.1 國內外研究現狀

文本情感傾向分析是一種交叉學科，涉及多個領域的知識和技術，特別是計算機科學、統計學、社會學和語言學等相關領域的知識，使一個能夠對意見、觀點、情緒和情感等內容進行計算和分析的研究領域，同時也被稱為情緒分類或意見挖掘。最近幾年，情感分析越來越多地被應用在各個領域中。

在基於文本的情感分析中，通常有三種主要方法：基於情感詞庫的情感分析、基於機器學習的情感分析與基於深度學習的情感分析。基於情感詞庫的方法使用預定義的情感詞匯數據庫進行情感分析，該方法具有高效、實用和易於實現的特點。楊廉正等人 (2022) 在原有的情感詞典基礎上，對 BiliBili 視頻網站上視頻評論的情感詞進行匯總和整理，構建出適用於 BiliBili 視頻評論領域的情感詞典，並運用 SO-PMI 算法進行其中的情感傾向計算，實現了一個面向 BiliBili 視頻評論的情感傾向分類系統。

基於機器學習的方法則是針對具體的情感分類問題，通過訓練機器學習算法來建立分類模型，並將未知文本數據分類至不同的情感類別中，是目前應用最廣泛的情感分析方法。常用的機器學習算法包括支持向量機、樸素貝葉斯、隨機森林等。陳波 (2021) 在其研究中設計並實現了一種基於機器學習的酒店用戶評論情感分析系統，將中文文本情感分析模型嵌入到情感分析系統的相關模塊中，滿足用戶的數據可視化與情感分析的需求。相比於基於情感詞庫的情感分析方法，該方法具有相對較高的準確性和魯棒性。

基於深度學習的情感分析技術使用了深度學習算法與神經網絡模型，通過學習大量真實數據來從文本中自動提取特徵並進行情感分析，能夠更精確地分析文本的情感特徵。其優點是準確性高、自適應性強和魯棒性強；缺點是需要大量的標注數據、可解釋性差和訓練時間長。因此在實際應用中需要綜合考慮多種因素來選擇具體的應用技術。在該領域，Ian Osband 等人 (2017) 討論了深度學習在情感分析中的優越性，如模型的自動特徵提取、模型的可擴展性和複雜性，以及模型在處理多語言情感分析中的潛力。

劉曉彤等 (2019) 基於傳統機器學習的方法，加入深度學習模塊，對在線評論進行情感分析與對比。對比機器學習模型 (MLP、SVM、樸素貝葉斯等) 和深度學習模型 (CNN、LSTM、BI-LSTM 等) ，比較實驗結果，提出優化方向。文浩宇 (2020) 結合了深度學習語言模型及傳統的機器學習模型，並在集成學習的思想上建立情感分析模型，以保證精確度又降低複雜度，用以分析餐飲行業的評論情感傾向。

目前來看，如何精確的進行文本情感傾向分析是一項複雜且龐大的任務，通過文本情感傾向分析，我們可以瞭解社會中不同群體和個人的情感狀態和趨勢，幫助我們更好地理解社會、政治、文化和商業等方面的問題，並為我們做出更加客觀、合理和有效的決策提供支持和幫助。

2.2 文本情感分類相關理論與技術

2.2.1 文本特徵提取

文本特徵提取就是從文本數據中抽取出有價值的信息，以方便進行一些重要的自然語言處理任務，如文本分類、情感分析和主題提取等。這些特徵可以包括單詞出現頻率、詞組統計、詞性標注、詞向量表示、以及字母級的 n-gram 等。文本特徵提取的主要目的是將文本數據轉換成機器可以理解的數字向量形式，從而利用機器學習和數據挖掘等技術，更好地分析和理解文本數據。

TF/IDF (Term Frequency - Inverse Document Frequency) 算法是一種文本特徵提取技術，是信息檢索領域非常重要的搜索詞重要性度量技術，用以衡量一個關鍵詞在一個文本中的重要性。TF/IDF 有關鍵詞提取、文本向量化、文本相似度計算、搜索引擎和文本摘要五項主要功能，這項技術的優點在於可以過濾常見或無關緊要的詞，保留更重要的信息，也可以找到給定文本中關鍵字的相關性，更有效的對文本進行分類、相似度計算和聚類分析。但這項技術也存在著缺點：不能處理次序和語法結構，不能完全捕捉語義信息，且僅適用於簡單的文本。

BERT (Bidirectional Encoder Representations from Transformers) 是由 Jacob Devlin 等人 (2018) 提出的一種語言模型，其基本思想是通過大規模的無監督學習來訓練一個通用的語言表示模型，隨後將該模型微調到特定的下游任務上。BERT 可以雙向處理上下文信息，在處理長文本、多義詞及歧義詞等方面具有較好的性能表現，但缺點是需要大量的計算資源和時間進行預訓練、標注數據與模型微調 (丁申宇, 2022)。

2.2.2 中文文本處理

中文文本處理是指對中文文本進行分析、處理、加工、應用的一系列技術和方法。它包括分詞、詞性標注、命名實體識別、句法分析、情感分

析、文本分類、文本生成等多個方面，主要是為了更加深入、全面地理解中文文本信息，以便更好的挖掘文本中所含信息。

有別于英文使用空格作為單詞之間的間隔，中文的詞與詞之間並沒有明顯的分割標誌，這會為文本特徵的提取帶來不小的挑戰。因此，中文分詞是中文文本處理的第一步，是指將一段中文文本按照一定的規則或算法，將其切分成一個個有意義的詞語。中文分詞可以採用基於詞典、基於統計等不同方法。其中基於詞典的方法通過字符串匹配的方式，能夠相對容易的識別文本中出現的中文詞語，並進行分割，但對詞典中尚未登記的詞語與一些多義詞的識別能力較弱，且匹配所需的計算量較大。基於統計的方法通過臨近字的共現評率確定字詞之間的關係，是目前主要的中文分詞實現思路，其優勢在於分詞能力不受限於詞典規則，能夠依賴統計算法實現字詞關係的挖掘，但缺點在於需要大量的語料庫進行模型訓練與測試，實現方式較為繁瑣，且分詞準確度會受到訓練文本的語境、質量等方面的影響。

從應用的角度出發，基於 Python 的中文 NLP 處理包 Jieba[2]是公認的優質中文分詞工具，其基本原理是通過 Trie 樹結構對文本進行掃描，並根據動態規劃算法查找最大概率的路徑，輸出分句中的詞組成相應的有向無環圖結構。同時，Jieba 分詞採用基於 HMM 模型使用 Viterbi 算法來解決未登錄詞的識別問題。在具體應用中，Jieba 分詞可以幫助我們有效地將中文文本切分成有意義的單詞，從而為後續的自然語言處理任務提供基礎支持。

2.2.3 文本分類算法：邏輯回歸和 Softmax

文本分類是指將一段文本分配到預定義的不同分類中的過程，通常是利用機器學習和自然語言處理技術完成的。一般可以使用的算法包括樸素貝葉斯、邏輯回歸、支持向量機、最大熵模型等，其中邏輯回歸是一種用於二類問題的監督學習算法，它通過對輸入特徵進行的線性加權和 sigmoid 函數的組合來計算輸出結果，輸出結果表示輸入數據屬某個類別的概率預測值。邏輯回歸的優點在於計算速度快、可解釋性好、可處理缺失值和可在線學習；缺點首先是對特徵空間的線性可分性要求很高，只適用於線性可分的問題，且容易受到噪聲和異常值的影響。同時，邏輯回歸能力有限，對於複雜的分類問題很難得到好的結果。

Softmax 算法可以被視作為邏輯回歸的擴展。其將每個輸出節點的分數通過指數函數轉換為概率分佈，並通過交叉熵損失函數進行優化。Bishop (2006) 介紹了 Softmax 的原理和應用，同時深入探討了 Softmax 在模式識

[2] https://github.com/fxsjy/jieba

別和機器學習中的重要性和應用場景。Softmax 非常容易實現、計算速度快，且可以適用於多分類問題。但是對於大規模數據集的處理速度就相對較慢，也很容易受到異常值的幹擾。

在後續實驗中，我們將基於 TF-IDF 算法與邏輯回歸算法，完成機器學習模型的建立，同時基於 BERT 與 Softmax 算法完成深度學習模型的建立，並對兩個模型在餐廳評論語境下的情感分類性能進行評估，選擇出效果較佳的模型。

3. 研究思路與實驗設計

3.1 數據集介紹與文本長度統計

3.1.1 數據集介紹

我們採用由美團發佈的 ASAP (Aspect category Sentiment Analysis and rating Prediction) 數據集 (Bu, et al., 2021)進行模型的訓練。這是一個大規模的中文餐廳評論數據集，用於進行評論方面類別 (Aspect category) 情感分析和評分預測。該數據集包含了來自大眾點評的 46,730 條餐廳評論，每條評論都手動標注了對 18 個細粒度方面類別的情感極性。數據集已被劃分為訓練集、驗證集與測試集，其中訓練集包含 36,850 條數據，驗證集與測試集各包含 4940 條數據。

ASAP 是目前最大的中文大規模評論數據集，可用於進行方面類別情感分析和評級預測任務。通過使用 ASAP 數據集，研究人員可以更好地理解用戶對餐廳的看法，並為企業提供更好的商業智能決策。

該數據集中包含了以下內容：

(1) 評論文本 (review)

該部分內容來自大眾點評網的用戶評論文本。考慮用戶隱私問題，評論所包含的用戶信息，如用戶名、頭像與發佈時間均不包含在數據集中。同時，少於 50 個漢字、超過 1000 漢字或非中文字符占比大於 70%的評論已被過濾。此外，美團還使用了自然語言處理技術和手動篩查相結合的方法，過濾低質量評論。評論樣本如附圖 2-1 所示。

(2) 評論評分 (star)

該部分來自用戶評價時給出的評分，共分為五個檔次 (1~5) 。根據實驗需要，之後我們會在數據預處理操作中，將此變量映射映射為-1 (負向情感) 、0 (中立情感) 與 1 (正向情感) 。

(3) 方面類別情感屬性

該部分為與用戶評論內容相關的方面類別及其情感屬性，共分為 5 個大類別、18 個細粒度方面類別，如附表 2-1 所示。每個方面類別包含-1、0、

1 三個情感極性，分別代表消極、中立與積極；對於評論中沒有提及的方面類別，標注為-2。

3.1.2 文本長度統計

使用 BERT 進行文本特徵提取前，需要統一文本的長度 (Yadav, Kumar, & Chauhan, 2020)，截取過長文本，並在 BERT 的文本預處理階段使用 [PAD] 標簽將短文本補長至目標長度。為了最大化保留文本內的信息，我們需要對數據集文本長度進行統計之後，選出合適的文本長度進行統一。圖 2-1 為該數據集文本長度的分佈圖。如圖所示，長度在 200 字~400 字之間的樣本占多數。考慮到性能問題，過長的文本可能導致訓練時佔用的顯存超過我們目前的硬件限制，導致訓練無法進行。結合 Google 給出的訓練建議[3]，我們將目標文本長度定在 256 字。

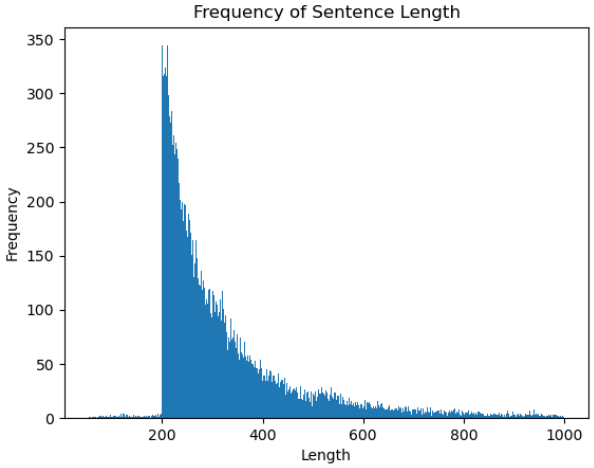


圖 2-1 評論文本長度分佈

3.2 模型與實驗設計

為了達到“客觀描述評論情感”的目的，我們在句情感模型的基礎上，引入細粒度方面類別情感模型，通過同時呈現評論文本整體情感與多個相關屬性的情感值，得到一個相對客觀的美食評論情感分析與預測方案， 從而更好地幫助消費者、店家與平臺對用戶評論進行分析，更好地做出決策。

在查閱資料的過程中，我們注意到，即使在大多數文本分類任務中，深度學習方法能夠達到更好的效果，但傳統機器學習方法仍能夠在某些場景下得到優於深度學習方法的結果。這主要取決於不同的文本特徵提取方法在不同類型的文本與文本分類任務中的表現。為了盡可能提高最終模型的可靠性，我們將嘗試基於傳統機器學習的方法 (TF-IDF 特徵提取+邏輯回歸) 與基於深度學習的方法 (BERT 微調模型) 進行模型訓練，並在比對

[3] https://github.com/google-research/bert#out-of-memory-issues

之後選用分類性能較好的模型作為我們的最終方案。

3.2.1 模型思路介紹

首先，我們基於 ASAP 數據集中的評論文本與用戶評分，建立美食評論文本的整體情感傾向模型，我們稱之為 S。在該模型中，文本特徵將作為自變量輸入模型，與因變量情感值建立回歸關係，得到一個二分類模型。

在確定了初步思路之後，我們注意到了用戶評分存在的主觀性：如附表 2-2 所示，同樣是關於"菜譜不錯，服務一般"的評論，兩位用戶給出了截然不同的評分；此外，我們也考慮到現實生活中存在的部分現象，如店家通過返利等措施誘導用戶給出好評，通過特定的程序故意、惡意刷評，或是顧客為了讓更多消費者注意到店家的不足之處，發佈差評內容的同時給出了較高的評分分數。以上現象存在較大的隨機性與不可控性，難以納入模型之中，往往會為模型帶來不可避免的誤差。因此，結合數據集中給出的方面類別情感值，我們決定在為評論文本整體情感進行建模的同時，為各個方面類別建立情感預測模型。

對方面類別情感進行建模時，我們嘗試了多種思路與建模方法，最終得到了一條較為合理的建模路線：先判斷各個方面類別在給定文本中的情感傾向是否明顯，再針對情感傾向明顯的方面類別分別進行情感值預測。對於前者，我們將分別為每個方面類別建立起情感顯著性的二分類模型，並稱其為 C_{1i}，其中 i 為方面類別的序號 $(i\epsilon[1,5])$。該模型中，文本特徵作為自變量，各個方面類別的情感傾向突出與否作為因變量。對於後者，我們將分別為每個方面類別建立起情感正負傾向的二分類模型，稱其為 C_{2i} $(i\epsilon[1,5])$，其中文本特徵作為自變量，各個方面類別的情感正負傾向作為因變量。在所有 C_{1i} 完成預測，給出文本中情感傾向顯著的方面類別之後，程序將單獨調用這些方面類別對應的模型 C_{2i}，分別預測情感正負值。

我們提出的最終思路如圖 2-2 所示。用戶輸入評論文本後，程序先進行文本特徵的提取與向量化，隨後將其分別輸入模型S與每個方面類別的情感顯著性模型 C_{1i}。在得到情感顯著的方面類別後，再調用對應的方面類別情感模型 C_{2i} 進行情感傾向預測，最終給出文本整體情感值、情感顯著的方面類別及其對應的情感傾向。

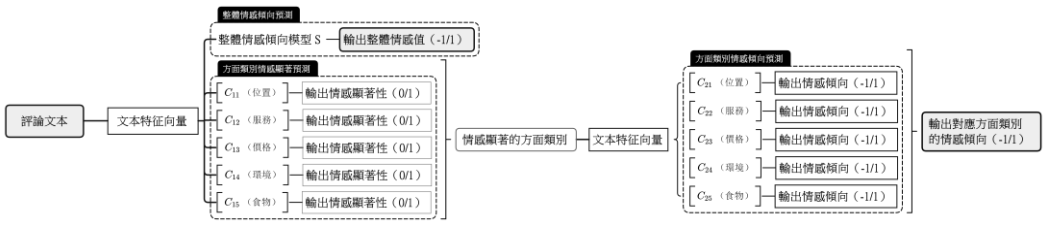


圖 2-2 模型思路與工作流

3.2.2 模型的選擇與算法解釋

基於回歸算法的思想，我們將分別嘗試使用傳統的機器學習方法與深度學習方法來完成建模。對於傳統機器學習方法，我們選擇基於 TF-IDF 算法的文本特徵提取與向量化，並使用邏輯回歸算法進行分類；對於深度學習方法，我們將使用評論文本對 Google 提供的 BERT 預訓練模型進行微調 (fine-turning) ，並完成其下游分類任務。最終，我們將對比上述各個方法中模型的 F1 值，從而確定在餐廳評論文本情感分析的場景下，預測效果最佳的模型。

(1) TF-IDF 特徵提取與邏輯回歸

TF-IDF (Term frequency-inverse document frequency) 是用於信息檢索領域的一種重要性搜索詞的度量技術，即對於某一詞匯對於文本整體所提供信息內容表達的重要性。TF 表示詞頻，即 term frequency，指關鍵詞 w 在總體文本D_i中出現的頻率，它的公式形式表達為：

$$TF_{w,D_i} = \frac{count(w)}{|D_i|} \qquad (2-1)$$

此表達式中，$count(w)$是 w 關鍵詞在文本中出現的頻率數，$|Di|$是文檔Di 中的單詞數，通過該比率數值得到詞頻。

逆文檔頻率 (IDF) 用於反應某詞匯的普遍程度，即當某一個詞普遍出現在文本中，大量文本包含該詞時，IDF 的數值較低，反之越高。IDF 的表達公式如下：

$$IDF_w = ln\frac{N}{1+\sum_{i=1}^{N} I(w,Di)} \qquad (2-2)$$

該表達式中，N 表示總文檔數目，$I(w,D_i)$ 為文檔 Di 中是否包括給定尋找的關鍵詞 w，當含有該詞時，i 為 1，反之為 0。 為避免搜索詞 w 不存在于所有文檔中帶來公式無法正常運算的問題，通過分母增加 1 進行平滑處理。根據以上兩定義式，搜索詞 w 在文檔 D_i 中的 TF-IDF 的表達如下：

$$TF-IDF_{w,Di} = TF_{w,Di} \times IDF_w \qquad (2-3)$$

由該表達式得出，當一個詞匯低普遍程度但高頻率出現在某文本中時，它的 TF-IDF 數值較高，反之越低。通過對搜索詞匯的 TF-IDF 計算可以得到所需語料中不同搜索詞的重要程度，為後期模型規劃建立基礎。TF-IDF 基於無監督學習原理，將詞頻與常規出現詞語的情況共同考慮，可以規避

普通詞匯，提供更多有效檢索詞匯。

TF-IDF 可以簡單快捷的處理所需文本，但使用詞頻分類文章內容的重要性不夠全面 (檢索詞出現次數不夠) ，故無法具體體現位置信息，也就無法檢索出相關重要性。

基礎的文本分類處理完成後，我們選擇使用解決分類問題的回歸算法：邏輯回歸，進行二分類的數據處理，輸出 0 或 1 的結果。邏輯回歸與線性回歸同屬廣義線性模型，且邏輯回歸中，設定因變量 y 服從二項分佈法則。

用於線性回歸的實數正態分佈是用均值參數化的。我們提供任何的均值都是有效的。二元變量上的分佈稍微複雜些，因為該模型的均值必須始終在 0 和 1 之間。解決這個問題的一種方法是使用 logistic sigmoid 函數將線性函數的輸出壓縮進區間 (0,1)。該值可以解釋為概率：

$$p(y = 1|x; \theta) = \sigma(\theta^{\top} x) \qquad (2-4)$$

在該方程中，p 是回歸模型通過 sigmoid 函數映射得到的概率，σ 為 sigmoid 函數。而 $\theta^{\top} x$ 是向量 θ 的轉置與 x 的點積，其值為各項 θ 與 x 相乘之和。

線性回歸中，我們能夠通過求解正規方程以找到最佳權重。相比而言，邏輯回歸會更困難些。其最佳權重沒有閉解。反之，我們必須最大化對數似然來搜索最優解。我們可以通過梯度下降算法最小化負對數似然來搜索。通過確定正確的輸入和輸出變量上的有參條件概率分佈族，相同的策 略基本上可以用於任何監督學習問題。

簡而言之，邏輯回歸是一種判別式模型，在線性回歸的基礎上，套用了一個 sigmod 函數，這個函數講線性結果映射到一個概率區間，並且概率在 0.5 周圍是光滑的，這就使得數據的分類結果都趨向於在 0, 1 這兩端。邏輯回歸原理的 sigmod 函數表達式如下：

$$g(z) = \frac{1}{1 + e^{-z}} \qquad (2-5)$$

由標準公式推導，本研究將使用以下表達式：

$$h_{\theta}(x) = g(\theta^{\top} x) = \frac{1}{1 + e^{-\theta^{\top} x}} \qquad (2-6)$$

$$g'(x) = \left(\frac{1}{1+e^{-x}}\right)' = \frac{e^{-x}}{\left(1+e^{-x}\right)^2}$$

$$= \frac{1}{1+e^{-x}} \cdot \frac{e^{-x}}{1+e^{-x}} = \frac{1}{1+e^{-x}} \cdot \left(1 - \frac{1}{1+e^{-x}}\right)$$

$$= g(x) \cdot \left(1 - g(x)\right) \qquad (2-7)$$

(2) BERT 與 Softmax

BERT 是基於雙向的 Transformer 編碼器實現的文本處理模型，其結構如圖 2-2 所示。而 Transformer 編碼單元是由 6 個 encoder 堆疊組成的，與其解碼層結構相同，其基本結構如附圖 2-3 所示 (李可悦, 陈轶, & 牛少彰, 2021)。

Bert 編碼器包含兩層，即 Self-Attention 及前饋神經網絡層，前者能夠協助當前節點獲取上下文語意。當神經網絡在所需大量語料的重複訓練達到足量輪數後，就能預測出相應詞的詞向量表徵，其原理類似於閱讀時通過上下文猜測中間句含義。兩層之間還包含一個 Attention 層，其協助當前研究頂點獲取需要關注的重點內容，其內部結構如附圖 2-4 所示：

a) BERT 模型的輸入與詞嵌入

BERT 模型的輸入通常為一對句子，並會在處理後轉換為一系列詞嵌入 (Word Embedding) ，包括 Token Embeddings, Segment Embeddings 及 Position Embedding。對於中文文本，我們首先需要進行分詞操作，並且每段文本以一個特殊的分類標簽 [CLS] 為首，通過特殊標簽 [SEP] 進行句分割。同時，過長的句子會在這一階段被截斷，過短的句子會使用特殊標簽 [PAD] 補全，其長度標準為模型的超參數之一，可根據樣本情況與訓練設備性能綜合考慮。接著，我們需要將文本轉換為上文提到的 3 類詞嵌入，其中 Token Embeddings 為詞本身的詞嵌入，通過查表將詞映射為對應的編碼；Segment Embeddings 以 [SEP] 標簽為標準，對不同句子進行區分；Position Embeddings 用於區別不同的詞所在的位置，解決了 Attention 機制中忽略語序的問題。最終以上 Embedding 進行相加，即可得到 BERT 編碼層的輸入。附圖 2-5 展示了以上過程，即 BERT 模型輸入的處理結果 (Devlin, Chang, Lee, & Toutanova, 2018)。

當存在使用 [PAD] 填充文本時，Attention Mask 將被引入，用於區分填充的無意義文本與真正有意義的文本部分 (Yadav, Kumar, & Chauhan, 2020)。

b) BERT 模型的預訓練過程

為了對語義表示能力的增強，BERT 模型的預訓練過程中包含 MLM (Masked Language Model) 與 NSP (Next Sentence Prediction) 兩個預訓練任

務，並使用兩部分損失之和進行模型權重更新與訓練：

i. MLM 任務

給定相應句子，隨機使用 [MASK] 標簽遮去一個或多個詞語，並對遮蓋詞進行預測，以達到完成文本深度雙向表示的目標。MLM 任務使得模型在遮蓋詞未知的情況下依賴上下文信息進行詞匯預測。

ii. NSP 任務

在給定的一對句子中，預測其在原文本中是否為相鄰語句，從而嘗試學習、理解兩個相連句子之間的關係。其中，第二個句子有 50%的可能是相鄰句，也有 50%的可能來自語料庫中的隨機句子。本質上 NSP 訓練是一種二分類模型，判斷前後句子之間是否與彼此相關。分類的結果將會包含在模型輸出的特殊向量 [CLS] 中。

在 BERT 模型的預訓練過程中，結合 NSP 與 MLM 任務的聯合學習能夠使模型既能學習 Token 的級別信息，同時能更加有效地的處理並理解語句和文本的語義信息，其損失函數如下：

$$L\left(\theta,\theta_1,\theta_2\right)=L_1\left(\theta,\theta_1\right)+L_2\left(\theta_1,\theta_2\right)$$
$$=-\sum_{i=1}^{M}logp\left(m=m_i\middle|\theta,\theta_1\right)-\sum_{j=1}^{N}logp\left(n=n_i\middle|\theta_1,\theta_2\right)\qquad(2-8)$$

其中，θ為編碼器部分的參數，θ_1為 MLM 任務中編碼器所接輸出層中的參數，θ_2為 NSP 任務中編碼器所接分類器的參數。在 MLM 任務中，M 為被 mask 的詞集合，且由於該任務可被視為給定詞典大小|V|上的一個多分類問題，$m_i\in[1,2,\dots,|V|]$；NSP 任務是一個二分類問題，因此$n_i\in[IsNext,NotNext]$。

完成模型的初步任務訓練後，BERT 進行文本向量的輸出，包括字符級別的向量與最左側的特殊向量[CLS]，如附圖 2-6 所示 (Devlin, Chang, Lee, & Toutanova, 2018)。

受算力與時間限制，本研究並沒有從頭訓練一個 BERT 模型，而是基於 BERT 預訓練模型 bert-chinese-base 進行微調，在 Transformer 輸出之上添加一個 Softmax 層實現進行文本情感分類。

c) Softmax 分類

Softmax 是邏輯回歸在多分類問題中的衍生，同樣屬廣義線性模型。其基本原理如下：

假設有訓練樣本集 $\{(x^1,y^1),(x^2,y^2),\dots,(x^m,y^m)\}$，其中$x^i\in\mathbb{R}^n$ 表示第 i個訓練樣本對應的短文本向量，維度為 n，共 m 個訓練樣本；$y^i\in$

$\{1,2,\ldots,k\}$ 表示第 i 個訓練樣本對應的類別，k 為類別個數。 給定測試輸入樣本 x，令模型分佈函數為條件概率 $p(y=j|x)$ ，即計算給定樣本 x 屬第 j 個類別的概率，其中出現頻率最高的當前樣本 x 所屬的類別，最終函數會輸出一個 k 維向量，每一維度向量表示樣本屬當前分類的概率，並且模型將 k 維向量的和做歸一化操作，即向量元素的和為 1。其表達式如下，並最終通過梯度下降法優化 (段丹丹, 唐加山, 温勇, & 袁克海, 2021)：

$$h_0(x^i)=\begin{bmatrix}p(y^i=1|x^i;\theta)\\p(y^i=2|x^i;\theta)\\\vdots\\p(y^i=k|x^i;\theta)\end{bmatrix}=\frac{1}{\sum_{j=1}^{k}e^{\theta_j^T x^i}}\begin{bmatrix}e^{\theta_1^T x^i}\\e^{\theta_2^T x^i}\\\vdots\\e^{\theta_j^T x^i}\end{bmatrix} \qquad (2-9)$$

本文在以上基礎上實現基於 BERT 的中文文本情感分類算法，經過模型訓練後，最終得到相應的實驗數據結果。

3.2.3 數據預處理方法

為了實現上述模型的建立，並完成評論情感的分析與預測，我們將對數據集內的評論、評分與方面類別情感屬性兩部分進行處理。

(1) 評論文本整體情感模型 *S*

對評論情感進行建模前的數據預處理流程如下：

a) 情感值劃分

在原數據集中，評論評分的範圍為 1~5 分。其中，3 分可以被視為一個情感中立評分，高於 3 分的評論可以被視為正向情感評論，低於 3 分的評論可以被視為負向情感評論。根據該規則，我們對原數據集中的評論評分進行映射，劃分出最終用於訓練模型的情感值：1~2 分映射為負面情緒 (-1) ，3 分映射為中立情緒 (0) ，4~5 分映射為積極情緒 (1) 。

b) 刪去中立情緒樣本

相較於表達正面與負面情緒的文本，代表中立情緒的特徵並不顯著，容易在訓練分類模型時引入誤差，因此我們僅保留表達正面與負面情緒的樣本，刪去表達中立情緒的樣本。

得到的數據與數據分佈如表 2-1 所示。通過觀察數據分佈，我們發現數據樣本存在一定的不均衡性。為了消除樣本不均衡可能帶來的誤差，我們將占比較少的負向情感 (-1) 進行過採樣，使兩類樣本的數量趨於平衡。

類別	樣本數量
正向情感 (1)	29132
負向情感 (-1)	2477

表 2-1 整體情感數據分佈

(2) 方面類別情感模型 C_{1i} 與 C_{2i}

在處理方面類別情感數據時，我們將對各個樣本與對應的所有方面類別進行處理。

考慮到模型訓練時間等條件限制，我們暫且將 18 個細粒度方面類別總結為“位置” (Location) 、“服務” (Service) 、“價格” (Price) 與“食物” (Food) 5 個方面類別，便於後續的模型訓練。首先，我們將沒有提及某方面類別的樣本 (-2) 與對某方面類別情感中立的樣本 (0) 視為情感不顯著，並統一將對應的方面類別標注為 0。隨後，我們將各個大類別下細粒度方面類別的標簽值進行求和，並將負數值映射至情感標簽‘-1’ (負向情感) ，正值映射至情感標簽‘1’ (正向情感) ，原情感值為 0 (情感不顯著) 的樣本保持不變，最終得到正、負、不顯著三類情感標簽。

接著，我們根據上述實驗方法，將以上數據集分為兩組數據，其中一組用於訓練方面類別情感顯著模型，另一組用於訓練方面類別情感值模型。對於前者，我們將-1 (負向情感) 與 1 (正向情感) 統一視為情感顯著，並將樣本中符合上述條件的方面類別標注為 1 (情感顯著) ；對於後者，我們將方面類別標簽為 0 (情感不顯著) 的樣本剔除，保留情感顯著 (-1 與 1) 的樣本。

根據上文提到的 5 類方面類別與對應處理方法，我們進一步將兩份數據各分為 5 組，如表 2-2 所示。此時對數據分佈進行觀察，我們發現 10 組數據中，皆存在樣本分佈不均衡的問題。因此，我們依次對其進行過採樣處理，使樣本分佈趨近均衡，並在之後的建模中對比過採樣與否所帶來的效果。

方面類別	類別	樣本數量	方面類別	類別	樣本數量
Location	情感不顯著 (0)	23647	Location	正向情感 (1)	12010
	情感顯著 (1)	13203		負向情感 (-1)	1193
Service	情感不顯著 (0)	16635	Service	正向情感 (1)	15165
	情感顯著 (1)	20215		負向情感 (-1)	5050
Price	情感不顯著 (0)	20363	Price	正向情感 (1)	11863
	情感顯著 (1)	16487		負向情感 (-1)	4624
Ambience	情感不顯著 (0)	18078	Ambience	正向情感 (1)	15591
	情感顯著 (1)	18772		負向情感 (-1)	3181
Food	情感不顯著 (0)	9148	Food	正向情感 (1)	24154
	情感顯著 (1)	27702		負向情感 (-1)	3548

表 2-2 方面類別情感數據分佈

3.2.4 實驗流程介紹

為了完成上述模型的建模工作，我們制定了一套具體的流程，如圖 2-3 所示。

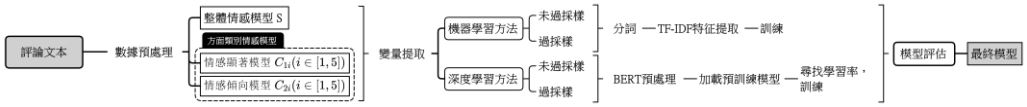


圖 2-3 建模流程

(1) 數據初步預處理

按照上述流程，處理各個模型的因變量，並對樣本占比不均衡的類別進行過採樣處理。

(2) 訓練評論文本整體情感模型 *S*

該階段，我們分別使用傳統機器學習方法與深度學習方法，提取文本特徵，並訓練兩組模型。

使用傳統機器學習方法建模時，我們使用中文 NLP 工具 jieba 完成文本分詞，借助 Scikit-Learn[4]中的 TF-IDF 向量化工具 (TfidfVectorizer) 完成文本特徵的提取與向量化，隨後使用 Scikit-Learn 建立邏輯回歸模型，並對模型性能進行評估。從初步實驗的結果來看，該方法在樣本分佈不均的情況下表現不佳。考慮到 TF-IDF 算法捕捉詞頻的原理，我們將分別使用原數據集與過採樣後的數據集進行 TF-IDF 特徵提取與邏輯回歸建模，並在之後評估不同模型的分類性能。

使用深度學習方法建模時，我們使用基於 TensorFlow 的 Keras 深度學習框架[5]與 ktrain[6]低代碼機器學習框架進行模型的訓練。Ktrain 已預先封裝好 Keras、Scikit-Learn、Hugging Face Transformers[7]等機器學習工具中的方法，包括基於 Keras 的 BERT 微調及下游文本分類任務，以及包括 jieba 在內的 NLP 工具，能夠高效地完成預處理、建模、訓練、模型評估與部署等操作。

在 ktrain 中，微調 BERT 預訓練模型並實現文本分類任務的流程如下：

a) 預處理

將文本按照 BERT 的分詞與標注規則進行處理，得到詞嵌入作為模型的輸入。這裡我們對訓練集與測試集中的評論文本進行處理。

b) 加載預訓練模型

[4] https://scikit-learn.org/

[5] https://scikit-learn.org/

[6] https://github.com/amaiya/ktrain

[7] https://huggingface.co/

這裡我們使用 BERT 中文預訓練模型 (bert-chinese-base) [8]，並將處理完畢的訓練集與測試集進行加載

c) 學習率尋找

Leslie N. Smith 曾提出過一種高效搜索合適學習率的方法，稱為循環學習率 (Cyclical Learning Rates) (Smith, 2017)。相較于通過大量實驗，逐步降低學習率的傳統搜索方法，這種方法會在訓練時，於合適的範圍內動態地、循環地調整學習率，從而快速找到合適的學習率用於正式訓練，提高模型訓練的效率與質量。Ktrain 提供了基於以上方法的學習率搜索工具，我們將借此確定合適的學習率，用於隨後的正式訓練。

d) 模型訓練與評估

經過以上步驟後，我們正式開始訓練模型。完成訓練後，我們保存模型，並評估模型性能。

(3) 訓練方面類別情感模型 C_{1i} 與 C_{2i}

該階段，我們的實驗方法與設計與上階段大體相同：同樣使用 TF-IDF 特徵提取+邏輯回歸與 BERT 兩種方法進行建模，但在數據的處理與模型的變量選擇上存在不同。在建立方面類別情感顯著與情感傾向模型時，我們將分別對五個方面類別進行建模，最終得到 10 組模型 ($C_{11} \sim C_{15}$, $C_{21} \sim C_{25}$)，並對每組中的使用不同建模方法的模型 (機器學習或深度學習、是否過採樣等) 進行評估，選出表現較好的模型作為該組的最終模型。

3.3 實驗結果和分析

3.3.1 實驗環境

機器配置：CPU 為 Intel Core i9-10900K，GPU 為 Nvidia GeForce RTX 3090，內存為 64 GB。

操作系統：Windows 11。

深度學習框架：TensorFlow，Keras 及 ktrain。

機器學習庫：Scikit-learn。

3.3.2 實驗結果展示

該部分中，我們將分別對使用不同數據處理方式與使用不同算法進行訓練的模型進行比較，並根據其分類性能，篩選出最終的模型S、模型C_{1i}與C_{2i}。由於模型 S 與模型 C_{1i}、C_{2i} 表現出的結果相似，這裡我們以模型 S 為例進行詳細介紹。

在模型評估中，我們使用了準確率、召回率、F1 值為指標來評估模型

[8] https://huggingface.co/bert-base-chinese

的性能。其中，準確率指模型正確預測的樣本占總樣本數的比例，精確率指模型預測為正樣本的樣本中，實際為正樣本的比例，如式 5-1 所示，其中 TP 代表 True Positive，FP 代表 False Positive；召回率指實際為正樣本的樣本中，被模型預測為正樣本的比例，如式 5-2 所示，其中 FN 代表 False Negative；F1 值的計算公式為式 5-3，式中 P 代表 Precision，R 代表 Recall，從中可以發現，其綜合考慮了精確率和召回率兩個指標，是一個綜合評估模型表現的指標，使模型的評估結果更具說服力。

$$Precision(P) = \frac{TP}{TP + FP} \qquad (2-10)$$

$$Recall(R) = \frac{TP}{TP + FN} \qquad (2-11)$$

$$F1 = \frac{2R}{P + R} \qquad (2-12)$$

同時，我們還將使用混淆矩陣對模型性能進行評估。混淆矩陣是用於評估分類模型性能的矩陣，它展示了模型對於真實類別的分類情況，其代表意義如附表 2-3 所示。

此外，我們將考慮模型的魯棒性，即模型在不同數據集和不同條件下的表現如何，是一個模型是否具有泛化能力的重要指標，以及模型的可解釋性，即模型是否可以解釋其預測結果的原因，是否能夠提供對預測結果的可信解釋。

(1) 不使用過採樣的 TF-IDF 模型和 BERT 模型對比

在本實驗中，我們採用未經過過採樣的數據進行訓練，並基於上文提到的不同算法分別訓練了對應的模型、比較其分類性能，最終選擇出該實驗中性能較好的模型。首先，我們基於機器學習方法進行整體情感模型的訓練，得到的模型評估結果如表 2-3 與表 2-4 所示：

模型	**precision**	**recall**	**f1-score**
-1	0.98	0.49	0.65
1	0.96	1.00	0.98

表 2-3：基於 TF-IDF 特徵提取與邏輯回歸模型的測試結果

	預測為-1	**預測為 1**
實際為-1	164 (TN)	174 (FP)
實際為 1	4 (FN)	3881 (TP)

表 2-4：基於 TF-IDF 特徵提取與邏輯回歸模型測試的混淆矩陣

其中，TN (True Negative) 表示真實為負類且被正確預測為負類的樣本數；FP (False Positive) 表示真實為負類但被錯誤地預測為正類的樣本數；FN (False Negative) 表示真實為正類但被錯誤地預測為負類的樣本數；TP (True Positive) 表示真實為正類且被正確預測為正類的樣本數。

從結果來看，在基於 TF-IDF 特徵提取與邏輯回歸的模型中，針對-1 類別，模型的精確率為 0.98，召回率為 0.49，f1-score 為 0.65，意味著模型對負面評論有一定的誤判可能性 (誤判率約為 0.51) ，但能夠正確分類的準確率很高；針對 1 類別，模型的精確率為 0.96，召回率為 1.00，f1-score 為 0.98，意味著模型在正面評論的判別上非常準確。而對於整體數據集，模型的準確率為 0.96，加權平均 f1-score 為 0.95，表現較好。綜合來看，模型對正面評論的判別準確度很高，而在負面評論的識別上還有一定的提升空間。

表 2-4 中，模型對於類別 1 的預測效果較好，True Positive (TP) 為 3881，False Negative (FN) 為 4，而對於類別-1 的預測效果相對較差，True Negative (TN) 為 164，False Positive (FP) 為 174。這意味著模型更容易將負類樣本誤分類為正類，但對於正類樣本的分類效果較為準確。因此，從魯棒性的角度來看，該模型需要改進。

模型	**precision**	**recall**	**f1-score**
-1	0.87	0.75	0.81
1	0.98	0.99	0.98

表 2-5：基於 BERT 模型的測試結果

如表 2-5 所示，在基於 BERT 訓練的整體情感模型中，對於類別-1，模型的精確率為 0.87，即模型在預測為負面評論時有 87%的準確率；召回率為 0.75，即模型能夠正確預測出 75%的負面評論；f1-score 為 0.81，表示模型綜合了精確率和召回率，該類別的分類表現較好。

對於類別 1，模型的精確率為 0.98，即模型在預測為正面評論時有 98%的準確率；召回率為 0.99，即模型能夠正確預測出 99%的正面評論；f1-score 為 0.98，表示模型綜合了精確率和召回率，該類別的分類表現較好。

	預測為-1	**預測為 1**
實際為-1	255 (TN)	83 (FP)
實際為 1	39 (FN)	3846 (TP)

表 2-6：基於 BERT 模型測試的混淆矩陣

從魯棒性的角度來看，混淆矩陣展示了模型對於不同類別樣本的分類

效果，即模型的魯棒性。從表 2-6 中可見，在本例中，模型對於類別 1 的預測效果較好，True Positive (TP) 為 3846，False Negative (FN) 為 39，而對於類別-1 的預測效果相對較差，True Negative (TN) 為 255，False Positive (FP) 為 83。這意味著模型更容易將負類樣本誤分類為正類，但對於正類樣本的分類效果較為準確。因此，從魯棒性的角度來看，該模型相對優於機器學習模型，但仍需要改進。

從混淆矩陣的角度來看，基於 BERT 模型的表現要優於基於 TF-IDF 特徵提取與邏輯回歸模型。基於 BERT 模型在預測負面情感的時候比基於 TF-IDF 特徵提取與邏輯回歸模型有更少的誤判，這可以通過比較兩個模型中負面情感的召回率來看出來。在基於 BERT 模型中，負面情感的召回率為 0.75，而在基於 TF-IDF 特徵提取與邏輯回歸模型中，負面情感的召回率為 0.49。同時，基於 BERT 模型中的準確率、精確率和 F1 得分也普遍都優於基於 TF-IDF 特徵提取與邏輯回歸模型。特別地，基於 BERT 的模型在負面評論的識別上表現得更好，準確率、召回率和 F1 值均高於傳統機器學習方法

(2) 使用過採樣的 TF-IDF 模型和 BERT 模型對比

在本實驗中，我們採用過採樣後的數據進行整體情感模型的訓練，並基於上文提到的不同算法分別訓練了對應的模型、比較其分類性能，最終選擇出該實驗中性能較好的模型。首先，我們基於機器學習方法進行模型訓練，得到的模型評估結果如表 2-7 所示：

模型	**precision**	**recall**	**f1-score**
-1	0.60	0.85	0.70
1	0.99	0.95	0.97

表 2-7：基於 TF-IDF 特徵提取與邏輯回歸與過採樣數據的測試結果

	預測為-1	**預測為 1**
實際為-1	288 (TN)	50 (FP)
實際為 1	193 (FN)	3692 (TP)

表 2-8：基於 TF-IDF 特徵提取與邏輯回歸與過採樣數據測試的混淆矩陣

從表 2-8 中可見，對於類別為-1 的情感，模型的準確率較低，只有 60%，但是召回率較高，達到了 85%。這意味著，分類器正確地預測了一部分-1 類別的樣本，但是也把一些屬 1 類別的樣本錯誤地預測為了-1 類別。綜合以上情況，該模型的 F1 分數為 0.7；對於類別為 1 的情感，模型表現非常好，準確率高達 99%，召回率為 95%，F1 分數也非常高，達到了 0.97，明顯優於負類的預測結果。

對於使用深度學習方法的模型，我們得到的評估結果如下：

模型	precision	recall	f1-score
-1	0.73	0.83	0.77
1	0.98	0.97	0.98

表 2-9：基於 BERT 模型與過採樣數據的測試結果

	預測為-1	預測為 1
實際為-1	279 (TN)	59 (FP)
實際為 1	104 (FN)	3781 (TP)

表 2-10：基於 BERT 模型與過採樣數據測試的混淆矩陣

從表 2-9 中可見，對於類別為-1 的情感，模型的精確度為 0.73，即在所有被預測為負面情感中，有 73%的預測是正確的；召回率為 0.83，即在所有實際為負面情感中，有 83%被正確地識別出來；F1 得分為 0.77，即模型在兩者之間取得了一個相對平衡的得分。對於類別為 1 的情感，模型的精確度為 0.98，即在所有被預測為正面情感中，有 98%的預測是正確的；召回率為 0.97，即在所有實際為正面情感中，有 97%被正確地識別出來；F1 得分為 0.98，依然優於負類的分類表現。

(3) 關於是否使用過採樣處理的對比

我們還注意到，在兩個機器學習模型中，使用過採樣方法能夠顯著提升模型的分類性能；在深度學習模型中，使用原數據集訓練的模型為四個模型中分類性能最佳的模型，其表現優於所有使用過採樣方法的模型。

從魯棒性的角度來看，使用過採樣的 BERT 模型相較於未使用過採樣的 BERT 模型，其在負樣本 (類別為-1) 的識別上有所提升，但在正樣本 (類別為 1) 的識別上有所下降。由於過採樣方法增加了負樣本數量，使得模型更容易識別負樣本，但同時也可能會導致模型對正樣本的識別效果下降。

使用 F1 值進行評估時，四個模型的 F1 值如表 2-11 所示：

模型	F1 (-1)	F1 (1)
原樣本/TF-IDF	0.65	0.98
過採樣/TF-IDF	0.70	0.97
原樣本/BERT	0.81	0.98
過採樣/BERT	0.77	0.98

表 2-11：不同模型的 F1 值對比

3.3.3 結果分析

綜上所述，基於 BERT 模型的表現最為出色，基於 TF-IDF 特徵提取與

邏輯回歸模型的表現相對較差；而過採樣方法對於模型的表現有一定的提升，但並不是一種通用的解決方法。尤其是在本文數據量有限並且算力有限的情況下，過採樣不僅會提高過擬合的概率，而且在訓練過程中，會對算力產生極大挑戰。

在這四種模型中，基於 TF-IDF 特徵提取與邏輯回歸的模型在正負樣本不平衡的情況下魯棒性較差，容易出現假陽性 (False Positive) 的情況；而基於 BERT 的模型相對來說魯棒性較好，能夠更好地區分正負樣本，表現出更高的精度和召回率。這主要是由於 BERT 在上下文語義理解方面的優勢使其能夠更好地提取文本特徵與語義信息，優於 TF-IDF 通過詞頻提取文本特徵的方式。因此，在餐廳評論語境下的文本整體情感分類任務中，使用基於 BERT 的深度學習模型能夠取得更好的效果。在基於 TF-IDF 特徵提取與邏輯回歸的模型中使用過採樣的方法可以提高魯棒性，但是仍然存在較高的假陽性率。在使用過採樣的基於 BERT 的模型中，魯棒性相對於不使用過採樣的模型有所提高，但仍存在少量的假陽性情況。通過查看模型的精度 (Accuracy) 、召回率 (Recall) 和精確率 (Precision) ，我們發現當模型的精度很高，但召回率和精確率較低時，就需要特別關注假陽性的情況。

使用以上實驗方法對方面類別情感顯著模型與情感傾向模型進行訓練與評估後，我們發現以上結論同樣適用於這兩個模型。根據實驗結果，我們選擇出以下模型，作為模型 S 與模型 C_{1i}、C_{2i} 的最佳模型：

模型	**是否過採樣**	**使用何種算法**
評論文本整體情感模型 S	否	BERT
方面類別情感顯著模型 C_{1i}	除了“Food”使用**過採樣數據**外，其餘方面類別均使用原數據	BERT
方面類別情感傾向模型 C_{2i}	否	BERT

表 2-12：最終模型的選擇

對於模型效果的評價，F1 值是一個較為合理的評估標準，但在現實生活中，人們往往更加關注負向情感的評論：商家能夠基於此對其服務進行改進，消費者能夠借此選擇服務、菜品更優質的店家。但受到評分主觀性的影響，單純地根據用戶評分篩選負面評論往往不夠全面、客觀。因此，有效地從評論中篩選出真正的負面評論顯得尤為重要。基於本文研究成果，商家與消費者可以高效準確地篩選出負向情感的評論，並從五個方面類別

出發，對評論內容進行更細緻的分析，從而進一步做出決策。圖 2-4 為最終供用戶使用版本的原型，其中包含評論文本、整體情感傾向、情感顯著的方面類別及其情感傾向。

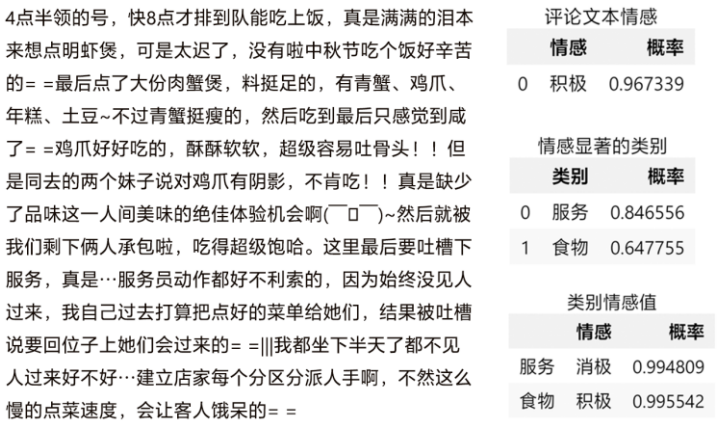


圖 2-4 最終輸出示意

4. 結論和展望

4.1 工作總結

本文基於自然語言處理和回歸算法，探究了餐廳評論的情感分析和預測問題。我們構建了一個具有良好泛化能力的深度學習模型，對評論數據集進行了建模和預測。

在實驗中，我們首先對數據集進行了探索性數據分析，發現了數據集的一些特徵和規律。接著，我們採用了預處理技術，包括分詞、情感值劃分等方法，對原始評論數據進行了處理，得到了可以輸入模型的特徵向量。然後，我們分別使用了多種回歸模型，對評論的情感得分進行了預測。最終，我們使用了模型評估指標，比較了不同模型的性能和準確度，並選出了可用、分類性能較佳的模型。

實驗結果表明，我們構建的深度學習模型可以對餐廳評論進行情感分析和預測，並且在測試集上取得了較好的表現。與傳統的回歸模型相比，我們的模型能夠更好地捕捉文本數據的分類特徵和複雜性。此外，我們的實驗還表明，在不同的語境與場景下，不同的模型可能表現出不同的性能，需要根據具體應用場景進行選擇和優化。

綜上所述，本文的實驗工作為餐廳評論情感分析和預測問題提供了一個有效的解決方案，具有一定的應用價值和推廣意義。未來，我們將進一步研究和改進該模型，以提高其預測能力和泛化性能。

4.2 不足之處

本文的研究存在以下幾個不足之處：

（1） 數據質量和分佈

由於評論門戶網站沒有提供相應的 API 可供獲取大量信息，同時也存在著驗證碼和反爬蟲的多種防火牆，限制了本文數據的獲取；而官方公開的數據獲取於 2021 年，比較缺乏時效性。這可能不足以覆蓋所有可能出現的情況，限制了模型的泛化能力。

（2） 本文中使用的模型相對簡單

可以探索更複雜的深度學習模型和更先進的自然語言處理技術，嘗試使用更複雜的下游模型 (CNN、LSTM 等) 與 BERT 組合以提高模型的準確性和魯棒性。同時，由於 BERT 模型對性能要求較高、速度較慢，可以嘗試 BERT 的蒸餾模型來提高訓練和推理速度

（3） 情感分析受到主觀因素的影響。

不同人對同一評論的情感分析可能不同。同時，同一個人對於不同店鋪，甚至在不同時期對於同一店鋪的評分標準也有可能不同。

（4） 本文沒有考慮到為了利益好評和惡意差評的特殊現象。

在實踐中存在大量抹黑或者因為利益驅使而給出好評的例子，這兩種現象都極大地影響到了感情的極性。在本文中，由於現有數據有限，無法通過更多維度展現水軍的。概率，從而優化模型並輔助商家和消費者判斷。

4.3 未來展望

為了進一步提高模型的準確性和泛化能力，我們將考慮以下方面的改進：

（1） 多維度數據的獲取以及更加詳細、準確的方面類別分類與標注。

可以考慮增加更多的數據維度，如菜品種類、服務質量、環境舒適度等因素，以更全面地分析用戶評論和情感。同時，可以考慮從不同來源獲取數據，如社交媒體、微博等，以增加數據的豐富性和多樣性。

（2） 建立水軍模型。

餐廳評論中存在大量的水軍評論，這些評論可能不真實，會對情感分析結果產生幹擾。可以考慮建立水軍評論識別模型，過濾掉這些不真實的評論。建立水軍模型。

從數據獲取的角度看，為了收集真實的用戶評論數據，可以採用一些有效的數據採集方法，例如抓取用戶在社交網絡上發佈的評論、從第三方數據供應商獲取用戶評論數據等等。這些數據採集方法可以獲得更加真實的用戶評論數據，從而降低水軍的影響。

從引入網絡安全技術的角度看, 為了避免水軍的影響，可以考慮建立多維度的水軍模型，包括用戶 IP、用戶登錄設備用戶。評論、歷史記錄等綜合建立模型，但由於目前獲取到的數據有限，後續可以通過其他能夠公開

數據的平臺或者是已有的公開水軍模型，分配普通用戶和水軍的權重，儘量保留二者特徵的同時減少水軍對整個模型的影響

（3） 探索更複雜的深度學習模型

使用更複雜的模型提高模型的準確性和魯棒性。同時，可以考慮進一步深入使用遷移學習等技術，從其他領域的數據中提取有用的特徵，加速模型的訓練與推理，提高模型的準確性。

綜上所述，未來的研究可以從多維度數據的獲取、建立水軍模型、探索更複雜的深度學習模型和提高模型的解釋性等方面展開，以進一步提高餐廳評論情感分析和預測的準確性和實用性。

5. 參考文獻

[1] Bishop, C. M., & Nasrabadi, N. M. (2006). Pattern recognition and machine learning. New York: springer.

[2] Bu, J., Ren, L., Zheng, S., Yang, Y., Wang, J., Zhang, F., & Wu, W. (2021). ASAP: A chinese review dataset towards aspect category sentiment analysis and rating prediction. *arXiv preprint arXiv:2103.06605.*

[3] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805.*

[4] Osband, I., Khandelwal, R., & Singer, Y. (2017). Advances in deep learning for sentiment analysis: A survey. *IEEE Transactions on Neural Networks and Learning Systems, 28(10)*, 2373-2385.

[5] Smith, L. N. (2017). Cyclical learning rates for training neural networks. *IEEE winter conference on applications of computer vision (WACV)* (pp. 464-472). IEEE.

[6] Yadav, J., Kumar, D., & Chauhan, D. (2020). Cyberbullying detection using pre-trained bert model. *International Conference on Electronics and Sustainable Communication Systems (ICESC)* (pp. 1096-1100). IEEE.

[7] 陳波. (2021). *基於機器學習的評論情感分析系統設計與實現.* 太原理工大學 .https://kns.cnki.net/KCMS/detail/detail.aspx?dbname=CMFD202201&filename=1021806156.nh

[8] 翟天智, & 楊廉正. (2022). 基於情感詞典的視頻評論情感傾向分析研究. *網絡安全技術與應用, (03)*, 53-56.

[9] 丁申宇. (2022). *基於 BERT-LDA 的在線評論細粒度情感分析.* 蘭州財經大學 .https://kns.cnki.net/KCMS/detail/detail.aspx?dbname=CMFD202301&filename=1022559696.nh

[10] 丁蔚. (2017). *基於詞典和機器學習組合的情感分析.* 西安郵電大學.

[11] 段丹丹, 唐加山, 溫勇, & 袁克海. (2021). 基於 BERT 模型的中文短文本分類算法. *計算機工程, 47(1)*, 79-86.

[12] 李可悅, 陳軼, & 牛少彰. (2021). 基於 BERT 的社交電商文本分類算法. *計算機科學, 48(2)*, 87-92.

[13] 劉曉彤, & 田大鋼. (2019). 融合深度學習與機器學習的在線評論情感分析. *軟件導刊(02)*, 1-4.

[14] 文浩宇. (2020). *面向餐飲評論的細粒度情感分析模型.* 中南財經政法大學 .https://kns.cnki.net/KCMS/detail/detail.aspx?dbname=CMFD202102&filename=1020748844.nh

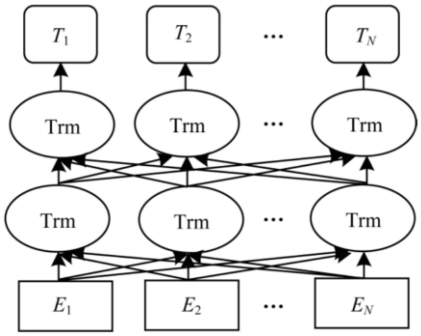
[15] 趙妍妍, 秦兵, & 劉挺. (2010). 文本情感分析. *軟件學報(08)*, 1834-1848.

6. 附錄

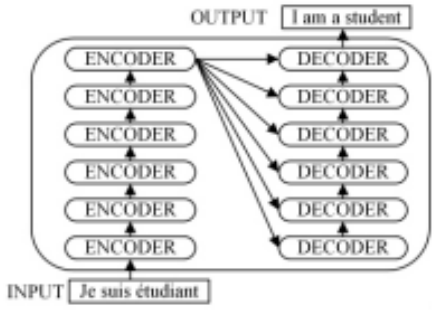
6.1 附圖

4 點半領的號，快 8 點才排到隊能吃上飯，真是滿滿的淚本來想點明蝦煲，可是太遲了，沒有啦中秋節吃個飯好辛苦的。最後點了大份肉蟹煲，料挺足的，有青蟹、雞爪、年糕、土豆~不過青蟹挺瘦的，然後吃到最後只感覺到鹹了雞爪好好吃的，酥酥軟軟，超級容易吐骨頭！！但是同去的兩個妹子說對雞爪有陰影，不肯吃！！真是缺少了品味這一人間美味的絕佳體驗機會啊！然後就被我們剩下倆人承包啦，吃得超級飽哈。這裡最後要吐槽下服務，真是…服務員動作都好不利索的，因為始終沒見人過來，我自己過去打算把點好的菜單給她們，結果被吐槽說要回位子上她們會過來的，我都坐下半天了都不見人過來好不好…建立店家每個分區分派人手啊，不然這麼慢的點菜速度，會讓客人餓呆的。

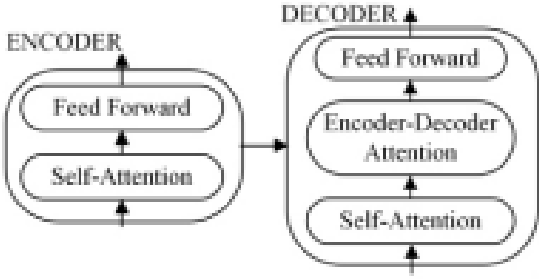
附圖 2-1 評論文本樣本示例



附圖 2-2 BERT 結構圖示



附圖 2-3 Transformer 結構圖示



附圖 2-4 Attention 層結構圖示

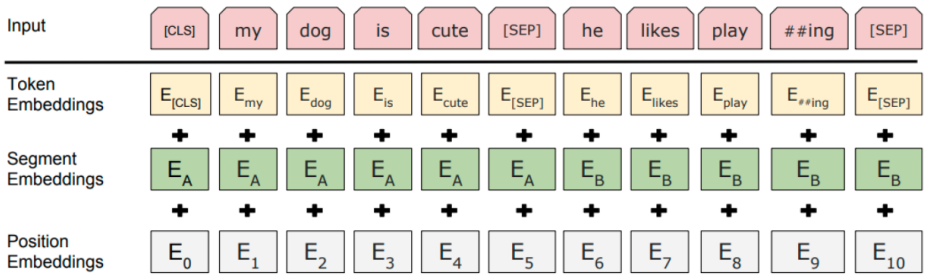
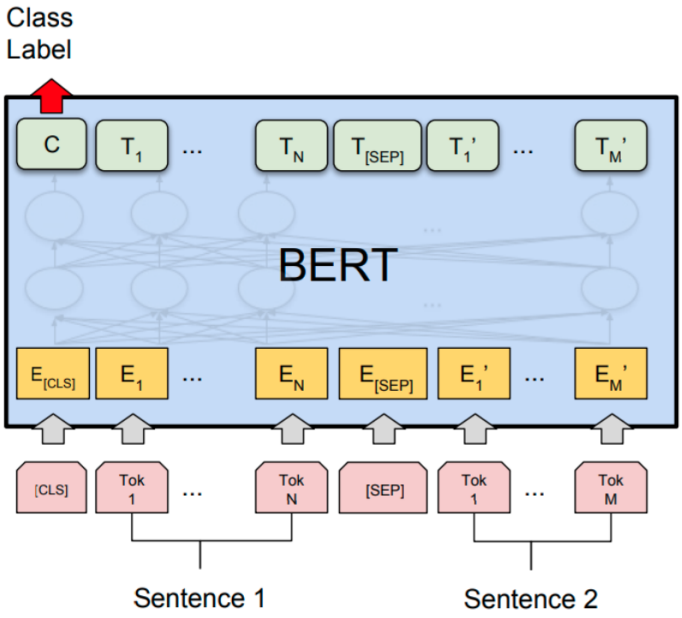


圖 2-5 BERT 的詞嵌入表示



附圖 2-6 BERT 模型輸出示意

6.2 附表

層次 1	層次 2
位置 (Location)	交通便利 (Transportation)
	商圈附近 (Downtown)
	容易尋找 (Easy_to_find)
服務 (Service)	排隊時間 (Queue)
	服務態度 (Hospitality)
	停車方便 (Parking)
	上菜速度 (Timely)
價格 (Price)	價格水平 (Level)
	性價比 (Cost_effective)
	折扣力度 (Discount)
環境 (Ambience)	裝修水平 (Decoration)
	嘈雜情況 (Noise)
	就餐空間 (Space)
	衛生情況 (Sanitary)
食物 (Food)	分量 (Portion)
	口感 (Taste)
	品相 (Appearance)
	推薦程度 (Recommend)

附表 2-1 方面類別展示

評論樣本	星級
●交通便利，在地鐵 1 號線客運中心站西北不遠處，b 出口最近。也是很多公交、長途汽車、機場大巴、車站接泊車的集中轉運中心。●單位年會，整體菜色擺盤尚可，口味一般過的去，價位不清楚。冷盤鳳爪挺好吃的，鹵香味，醃制的很清爽。瓜子有很香的奶香味，味道不錯。魚幹很酥，但是味道一般。酸辣大白菜還不錯，開胃清爽。葷菜沒有特別有印象的。●缺點:服務意識太差。冷菜上桌後，因為年會節目影響，等了很久才上主菜，在這段空隙時間服務員就一直站著也不收拾下空盤。等上主菜了，讓撤個虎碟都不樂意，更別說換了。也沒有第一時間對客戶要求的反饋，說個“稍等一下”也就一秒的事情。要酒杯要醋什麼的也要反復要求幾遍真心累。上菜也有點手忙腳亂的。	2 顆星
1.在萬國裡面算最多人的一家了，排隊足足排了兩小時，還有有微信查號和大眾點評排號的功能，這一點做得不錯。2.點了豆花烤魚，烤牛蛙，烤糖醋骨，配菜四樣，柑橘檸檬和香茅特飲，白麵條，一共 280.5，過得去吧。烤魚比較嫩，沒想像中辣。牛蛙好好吃，好香，肉很嫩。其他的東西就不過不失吧。發點評送白涼粉，好好吃。3.服務真的有點跟不上，上菜有點亂，比我們晚來的人，菜上得比我們快多了，讓服務員跟跟單，還十萬個不願意。加水什麼也沒人搭理。4 總體而言，值得嘗試，但需要提高服務員的培訓和素質，畢業這是品牌的生招牌。	5 顆星

附表 2-2 評論評分受主觀性影響

類別	解釋
TN (True Negative)	真實為負類且被正確預測為負類的樣本數
FP (False Positive)	真實為負類但被錯誤地預測為正類的樣本數
FN (False Negative)	真實為正類但被錯誤地預測為負類的樣本數
TP (True Positive)	表示真實為正類且被正確預測為正類的樣本數

附表 2-3 混淆矩陣意義解釋

7. 致謝

在此，我們感謝所有在本文撰寫期間給予我們幫助與支持的人們。他們的貢獻、指導與鼓勵，使我們能夠順利地完成這項研究，並取得了一定的成果。

首先，感謝我們的指導老師姜慧敏教授。她對我們的論文進行了悉心的指導，提出了許多中肯的建議，在我們的研究過程中給予了無私的支持與幫助。她的專業水平、嚴謹態度、敬業精神，對我們產生了深遠的影響。

同時，感謝我們小組內的每一位成員。本文的實驗與撰寫過程中，離不開大家的相互溝通、幫助與支持。每一位成員都在自己的領域內做出了積極的貢獻，展現了良好的團隊合作精神。我們互相學習、共同進步，一同克服了重重困難，順利完成了本次的研究。

此外，本文所使用的開源軟件也為我們的研究提供了極大的便利與支持。在此，我們感謝所有開源軟件開發者的付出與他們無私奉獻的精神。他們為科學研究提供了強大的工具與平臺，促進了知識的傳播與創新。

最後，感謝我們的家人與朋友對我們的支持與鼓勵。您們的關心、理解與支持，是我們完成這篇論文最堅實的後盾。