元智大學
資訊管理學系碩士班
碩士論文

基於辭典的意圖探勘方法及其在電影票房預測的應用

A Lexicon-based Approach for Intention Mining and Its Application in Movie Box Office Prediction

研 究 生：鄭鈺均

指導教授：楊錦生 博士

中 華 民 國 一一二年七月

基於辭典的意圖探勘方法及其在電影票房預測的應用

A Lexicon-based Approach for Intention Mining and Its Application in Movie Box Office Prediction

研 究 生 ： 鄭鈺均　　Student： Yu-Chun Cheng

指導教授 ： 楊錦生 博士　　Advisor： Dr. Chin-Sheng Yang

元智大學

資訊管理學系碩士班

碩士論文

A Thesis

Submitted to Department of Information Management

College of Informatics

Yuan Ze University

in Partial Fulfillment of the Requirements

for the Degree of

Master of Science

in

Information Management

July 2023

Chungli, Taiwan, Republic of China.

中華民國一一二年七月

論文口試委員審定書

基於辭典的意圖探勘方法及其在電影票房預測的應用

學生：鄭鈺均　　指導教授：楊錦生 博士

元智大學資訊管理學系碩士班

摘要

隨著網路的普及和社群媒體的興起，人們獲取資訊的方式逐漸改變。社群媒體賦予使用者表達觀點和創作內容的自由。商業公司則透過蒐集和分析使用者在網路上的評論資料，深入了解消費者對產品的感知和喜好。對其他使用者而言，這些評論內容是影響他們是否有動機消費的重要參考，因此判斷評論有無意圖也就成為關鍵的因素。本研究探討使用者在電影平台 IMDb 上的評論是否具有意圖，提出了使用 Word2vec 詞嵌入方式來擴充同義詞。與傳統的字典法不同，Word2vec 能生成包含多個單詞的相似詞彙。該模型基於大量文本訓練，具有實時性且能捕捉熱門趨勢詞，從而更全面地考慮使用者可能使用的詞彙。透過挑選意圖種子字，在 Bigram 訓練文本的 Word2vec 模型中擴充意圖字，並根據擴充次數建立不同的意圖辭典，產生意圖特徵。最後，使用支援向量回歸、線性回歸、決策樹和隨機森林等四種回歸型的機器學習方法，對所有與電影相關的變數進行分析，評估意圖辭典和機器學習方法對電影票房的影響。

關鍵字 ： 社群媒體、意圖探勘、Word2vec 詞嵌入、機器學習、電影評論

A Lexicon-based Approach for Intention Mining and Its Application in Movie Box Office Prediction

Student : Yu-Chun Cheng Advisor : Dr. Chin-Sheng Yang

Department of Information Management

College of Informatics

Yuan Ze University

ABSTRACT

With the internet and social media, accessing information has changed. Social media allows users to freely express viewpoints and create content. Companies collect and analyze user-generated comments to understand consumer perceptions. Comments influence others' purchase motivation, making intent analysis crucial. This study examines intent in user comments on IMDb using Word2vec to expand synonymous terms. Unlike dictionaries, Word2vec generates similar multi-word terms. Trained on a large corpus, it captures trending words and considers user vocabulary comprehensively. Intent seed words expand in the Word2vec model, creating different intent dictionaries based on frequency. Regression-based machine learning methods (Support Vector Regression, Linear Regression, Decision Tree, and Random Forest) analyze movie variables, assessing the impact of intent dictionaries and machine learning on box office performance.

Keywords: Social Media, Intention Mining, Word2vec Word Embedding, Machine Learning, Movie Reviews

誌謝

過去五年在元智學習到許多知識，結交各方好友，也接受了很多師長的教誨，我很榮幸也相當珍惜如此的學習環境和機會。過去大學的課程訓練奠定了我對於專業領域的認知，因為對於環境的熟悉，這讓我很順利的銜接到了碩士學程。在研究生生涯中最需要感謝我的指導老師楊錦生教授，除了每周的開會討論、相互指教想法，在遇到問題的時候老師總是會提出非常具體且明確的解釋，在撰寫論文的時候也會一步步指導並給予相當受用的建議，相當感謝老師這段時間的幫助以及指引，讓我在這當中學習到許多專業知識以及對待事情的態度和解決問題的能力。

感謝同學們一路上的相互扶持，謝謝實驗室的學姊云瑄，總是不藏私的大方分享自己所知道的事情，並給予可靠的意見，感謝實驗室的宣輔、紹偉，過去常常一起討論報告、聊天、吃飯，也搭了你們不少次便車，真的替我省了很多時間!謝謝楊毅、敬萱、唯穎，當初在就業博覽會的時候一起討論了許多求職辛苦談，希望大家實驗順利也找到好工作。謝謝佳璇在最後一個學期的時候還願意一起去上瑜珈課，也讓我們研究生活有喘息的時間。謝謝君庭，從大一認識你到現在你總是個很善良有趣的人，很關心朋友也是個很棒的傾聽對象，願你未來的研究順利。謝謝老同學亮妤，很開心過去這段時間互相分享生活上的瑣事，相互砥礪玩耍，希望你在工作上可以更加順心。最後感謝我的家人爸爸媽媽妹妹外公外婆，你們總是關心我的狀態，準備好吃的食物溫暖我的胃，謝謝你們的支持與付出我才能夠完成今天的學業，希望大家平安健康。

鄭鈺均 謹誌

中華民國一一二年七月

目錄

表目錄

圖目錄

第一章 緒論

1.1 研究背景

近年來，隨著網路和行動裝置的普及，人與人之間的資訊傳遞成本和時間大幅減少。科技的進步，促使社群媒體的出現，改變了人類接收資訊的方式，並逐漸取代了傳統媒體形式，例如: 新聞、報紙、雜誌、廣播、電視和電影等。社群媒體的多樣性，包含: 社群網絡、部落格、影音平台、即時通訊工具等，讓使用者擁有更多權力，可以自由撰寫或創作內容。這樣的互動模式讓人們可以更直接地表達自己的觀點、情感和想法，並與其他人進行互動。社群媒體的全球性使得內容可以輕易地傳播到世界各地，增進了內容的共享和全球的互聯。

Song et al. (2019)曾在文章中提過，過去十年間用戶生成內容[1] (User-generated content, UGC)或口碑[2] (Word-of-Mouth, WOM) 的參考價值也隨著使用者數量的增加而持續提升。這些由消費者自發性產生的資訊，具有更高的可信度，也較能夠反映消費者真實的需求和想法。因此，許多公司透過擴大其在社群媒體上的曝光度，利用 UGC 和 WOM 來促進產品或品牌的銷售，而這些社群媒體上的活動，也成為了公司行銷策略中不可或缺的一環。此外，對於電影等體驗類型的產品來說，由於其具有較強的感性屬性和情感性，因此 UGC 和 WOM 更是扮演了重要的角色。消費者會透過觀看其他人的評價和分享，來獲取更多關於電影的資訊和體驗，從而影響其最終的消費決策(Maslowska, Malthouse, and Viswanathan 2017)。社群媒體吸引了大量使用者，因此商業公司對於蒐集網頁或應用程式的流量數據已成為不可或缺的一環，透過分析社群媒體上的數據，能夠瞭解使用者對產品的需求，同時也針對潛在客戶提供更具吸引力的產品服務。(Zeng et al. 2010)。

意圖探勘[3]（Intention Mining）為一種探究使用者意圖的技術，透過分析使用者行為和語言等方面來獲取和推斷使用者的意圖。在社群媒體分析方面，可以應用在使用者評論和社群媒體貼文中，分析使用者對於產品或服務的意見和需求。透過意圖探勘分析線上購物者的經驗數據，可以預測客戶在瀏覽網路商店的網站時是否

[1] https://en.wikipedia.org/wiki/User-generated_content
[2] https://en.wikipedia.org/wiki/Word_of_mouth
[3] https://en.wikipedia.org/wiki/Intention_mining

具有購買意願，從而提高產品和服務的競爭力和滿意度(Kabir, Ashraf, and Ajwad 2019)。綜合上述，意圖探勘的商業應用可以為產品開發提供更有價值的洞見和建議，進而增加企業的收益和發展空間。

1.2 研究目的

意圖探勘的方法可以分為兩種: 機器學習法(Machine Learning Approach)以及辭典法(Lexicon-based Approach) (Dhaoui, Webster, and Tan 2017; Khattak et al. 2021)，在過去的研究中 Zhang et al. (2021)以 Airbnb 為例分析平台中使用者的評論內容，再進一步討論共享經濟中消費者感知價值與重複購買意願之間的關係。應用於電影上的案例有 Liu et al. (2016) 認為購買意圖探勘為一個二元分類問題，並提出關於購買意圖探勘的六個特徵類別，最後使用線性與非線性模型預測票房收入。另一個關於電影的應用是 Ahmad, Bakar, and Yaakub (2020)提出了辭典法來擴充與電影相關的意圖字，並進一步分析 Youtube 平台上電影預告片評論是否具有意圖。

本研究以 Ahmad, Bakar, and Yaakub (2020)做為主要研究參考論文，論文中關於意圖探勘辭典的部分使用字典進行擴充，意思是使用字典去搜尋相關替代的意圖字。然而這個方法存在著一些缺點:

- 只侷限於單一單詞的替換:
  在論文所提及的同義詞網站 thesaurus.com[4]中輸入單字查詢，但發現只能針對一個單字做同義詞的推薦。例如: 想要查詢“must buy”的同義詞，但網站無法輸入 Bigram 的詞，只能針對一個單字“must”或是“buy”做替換。

- 字典內容缺乏即時性:
  字典可能是幾年才會重新編寫或是收錄新單字，因此無法從字典中得到時下的新字詞。

[4] https://www.thesaurus.com/

基於上述兩項原因，因此提出以 Word2vec (Mikolov et al. 2013)詞嵌入方法來擴充意圖字。Word2vec 詞嵌入在許多文獻中都可以被用來當作同意詞擴充(Zhang, Li, and Wang 2017)，此方法有以下幾個優點:

- 可以產生 N-gram[5]的同義字:
  與傳統的字典法不同，Word2vec 不侷限於單個單字的同義詞生成。透過模型的訓練，能夠生成多個單詞的相似詞彙，包括: Unigram、Bigram、Trigram 等等。這種能力使得 Word2Vec 模型在同義詞生成方面可以捕捉到更多詞彙之間的關聯性，並且提供更具彈性和豐富的語言表達及理解。

- 生成的同義字包含熱門趨勢詞:
  Word2Vec 詞嵌入模型是基於大量文本訓練而成的，因此只需使用新文章作為訓練文本，無需等待人工編輯或擴充詞典，使其具有實時性。這種模型的優勢在於能夠捕捉熱門趨勢詞，並生成具有相似含義的詞彙。

- 可以依照專業領域，訓練不同的 Domain Model:
  透過將相關領域的文本資料作為訓練資料，Word2Vec 模型可以學習到該專業領域中詞彙的含義和上下文關係。這樣的訓練過程可以提高模型在特定領域的表現，使其能夠更好地理解該領域的專業術語和用語。例如: 在醫學領域，可以使用大量的醫學文獻和專業術語來訓練醫學領域的 Word2Vec 模型，以便在醫學文本分析或醫療診斷等任務中獲得更準確和專業的結果。

本研究探討在意圖探勘中，如何有效地擴充意圖字以提升意圖特徵的準確性和有效性。為此，提出了一個新的擴充意圖字的方法，透過自行訓練一個 Word2vec 電影相關的詞嵌入模型，並針對與電影相關的意圖詞進行擴充。透過這樣的方式，可以更全面地考慮使用者可能使用的詞彙，進而生成更具代表性的意圖特徵。

[5] https://en.wikipedia.org/wiki/N-gram_language_model

1.3 研究架構

本研究分為五個章節討論，第一章為緒論，介紹研究的背景、研究目的以及研究架構；第二章為文獻探討，本章節將分成兩個部分，回顧過去的意圖探勘技術方法和商業應用中的使用者意圖探勘研究案例；第三章為研究方法，將介紹與本研究模型相關的電影變數、資料前處理、Word2vec 意圖辭典的作法、機器學習方法、評估指標；第四章為實驗結果，介紹研究使用資料集以及各個實驗變數的敘述性統計，並詳細說明各個實驗結果和敏感度分析；第五章是實驗總結與未來展望，最後針對本研究分法及結果提出改善方針。

第二章 文獻探討

使用者意圖為一種心理狀態，讓人可以從中挖掘使用者的期望、需求及興趣(Habib et al. 2018)。例如，“我正在觀望最近的院線電影，想要周末去觀看”，以上可以看出消費者的意圖，同時消費者也有可能在尋找電影的相關資訊。根據 Rashid et al. (2021)的系統性文獻回顧，意圖可分為八個類別，包括：購買意圖（Purchase Intention）、行為意圖（Behavior Intention）、人類隱含意圖（Human Implicit Intention）、搜尋意圖（Search Intention）、持續意圖（Continuance Intention）、行動裝置使用意圖（Mobile Usage Intention）、查詢意圖（Query Intention）和一般意圖（General Intention）。另外此篇論文也有提到關於意圖探勘的六種資料類型，分別是:搜尋引擎日誌數據(Search Engine Log Data)、基於模型的生成數據(Model-based Generated Data)、問卷調查法(Questionnaire Survey Method)、通用數據集(Generic Datasets)、行動裝置使用數據集(Mobile Usage Dataset)、社群媒體數據集(Social Media Dataset)。鑑於本研究的目的是在社群媒體中文字評論的意圖探勘，因此將著重於用戶在社群媒體平台的購買意圖以及行為意圖方面的文獻。

2.1 意圖探勘方法

關於意圖探勘的文本分析可以分成兩個主要的技術層，分別為辭典法(Lexicon-based Approach)與機器學習法(Machine Learning Approach) (Dhaoui, Webster, and Tan 2017; Khattak et al. 2021)。辭典法首先建構一個包含已知相關意圖詞的辭典。在確認完成意圖辭典後，對於目標文本，可以根據辭典的內容進行偵測。如果文本中包含辭典中的意圖詞，則可以將其視為具有意圖(Hamroun and Gouider 2020)。Zhang et al. (2021)在研究中僅提出了簡單的三個詞(recommended, next time, again)，作為研究討論顧客再購以及推薦的意圖辭典。Symeonidis, Peikos, and Arampatzis (2022) 研究發現，相較於單獨出現的單詞，助動詞的加入可以更整的表達顧客意圖。因此，在文本資料分析中，研究人員提出了一個包含 20 個出現頻率最高的動詞片語以及一般單詞的辭典(looking, get, buy, want, need, purchase, use, replace, choose, take, find, like thinking, would like, should buy, would choose, could get, could find, would use, would have)，這個辭典可用於分析文本數據，以辨識並捕捉顧客的意圖。Ahmad et al. (2020)挑選出五個意圖詞(must buy, cannot wait, looking

forward, keep an eye on, must have)作為研究中的意圖種子字，透過 therasus.com 線上字典進行同義詞擴充，並建構購買意圖辭典用判斷文本是否具有購買意圖。

機器學習法會透過大量的文本訓練一個模型，並自動判別文本是否具有意圖，其文獻大多為監督式的機器學習法。Chen et al. (2013)提出了一種基於 Expectation Maximization[6] (EM)演算法的特徵選擇方法 FS-EM (Feature Selection Expectation Maximization)，以及基於遷移學習[7]的 Co-Class (Co-Classification)算法，用於論壇貼文的議題分類。目的是聚焦於提高意圖分類的準確性，並採用了特徵選擇和遷移學習的技術，以更好的調整模型辨別未知類別的文章。Ghiassi, Lio, and Moon (2015)的研究引入了基於卷積神經網路（CNN）的 CIMM (Consumption Intention Mining Model)模型，用於探勘使用者的消費意圖。此研究模型專注於解決遷移學習情境下的問題，通過使用深度學習方法和特徵學習，能夠更好的理解和分析隱性意圖，提升意圖判別的效能。Li, Du, and Wang (2017)提出了一種 Attention-Based RNN Encoder-Decoder 模型，專注於隱性意圖的偵測和意圖分類。通過結合 RNN 和 CNN 的方法，能夠將包含隱性意圖的文章轉換為顯性，從而實現更準確的意圖分類。Ding et al. (2018)的研究提出了一個深度遷移學習框架，用於消費意圖識別。此方法利用 Tree-LSTM 學習句子的樹結構特徵表示，並通過遷移學習技術將特徵表示層的參數從源領域轉移到目標領域，從而改善意圖識別的準確性。這些論文的研究結果突顯了機器學習和深度學習在意圖探勘中的應用價值，它們提供了不同的方法和技術，包含: 特徵選擇、遷移學習等等，以增強意圖辨識的效能和準確性。

辭典法和機器學習法是意圖探勘的兩個主要技術。辭典法利用建立意圖詞辭典的方式進行意圖偵測，而機器學習法則透過大量文本，訓練模型自動判別意圖。以往的論文大多是以機器學習法的意圖探勘為主，相較於此，辭典法的相關研究並沒有太多的參考文獻。

[6] https://zh.wikipedia.org/zh-tw/%E6%9C%80%E5%A4%A7%E6%9C%9F%E6%9C%9B%E7%AE%97%E6%B3%95
[7] https://zh.wikipedia.org/zh-tw/%E8%BF%81%E7%A7%BB%E5%AD%A6%E4%B9%A0

2.2 使用者意圖探勘於商業應用中的研究案例

關於使用者意圖探勘相關的商業應用文章數量其實並不多。Zhang et al. (2021)以Airbnb住宿租賃平台為例,提出了P2P[8]共享經濟中消費者感知價值的架構框架。該框架包括: 功能價值、享樂價值、認知價值和社會關係價值等構面。研究中使用了 LDA[9]主題模型（Latent Dirichlet Allocation）和情感分析方法來構建基於線上評論的感知價值測量指標,並應用辭典法從評論中提取了重複購買意圖變數。接著,使用結構方程模型來檢驗感知價值對重複購買意願的影響，最後發現社會關係價值(服務態度、主客之間的關係)為最重要的影響因素。

另一方面 Liu et al. (2016)針對電影票房的預測，抓取社群平台微博(Weibo)、中國電影網站 Wangpiao 作為使用者評論以及電影元數據的資料集，並研究提出了六個關於購買意圖探勘的特徵類別，包含: Bag-of-Word、是否提及其他使用者、是否含有網址連結、是否含有表情符號、文字長度是否超過 30 字元、是否含有意圖字，研究中過濾了 42 個字作為意圖觸發詞。一樣為預測電影票房的研究 Ahmad, Bakar, and Yaakub (2020)以 Youtube 電影預告片下方評論以及 Box Office Mojo 電影資料庫網站作為使用者評論與電影元數據的資料出處，討論如何從影評中準確提取電影購買意圖、並針對文本進行情感分析，研究使用辭典法，先對評論文本進行 TF-IDF[10]，挑出前 200 個重要的 Bigram 詞組，再利用人工判斷這些詞是否具有購買意圖,將過濾後的詞作為意圖種子字,並使用線上字典 therasus.com 進行擴充,建構電影評論購買意圖辭典 (Movie Reviews Purchase Intention Lexicon, MRPIL)，最後意圖偵測,若評論中含有一個或多個意圖辭典的詞，則判斷該評論具有意圖。

這些研究案例呈現了在商業領域中運用使用者意圖探勘的範例，並提供了一些方法和技術,用於從文字評論中擷取意圖相關信息。儘管目前相關文章的數量有限，這些研究仍然為商業應用領域在進一步探索使用者意圖探勘提供了啟發和參考，並為未來的發展提供了有價值的指引。

[8] https://zh.wikipedia.org/zh-tw/%E5%B0%8D%E7%AD%89%E7%B6%B2%E8%B7%AF
[9] https://zh.wikipedia.org/zh-tw/%E4%B8%BB%E9%A2%98%E6%A8%A1%E5%9E%8B
[10] https://zh.wikipedia.org/zh-tw/Tf-idf

第三章 研究方法

本研究的目標是在使用者評論意圖分析和票房預測方面探索文字探勘的應用。本章將詳細介紹實驗流程，包括資料來源、變數、資料前處理、Word2vec 模型學習、意圖辭典擴充、意圖分析、情感分析、模型建構和預測評估。

參照圖 1 研究流程圖，研究使用的資料集來自 IMDb[11] (Internet Movie Database) 網路電影資料庫網站，透過網路爬蟲收集了相關的電影資料，其中包括電影訓練資料集和用於 Word2Vec 模型訓練的電影評論資料。電影訓練資料集包含電影的元資料以及與這些元資料相關聯的電影評論。在完成資料蒐集後，從電影訓練資料集中提取了所需的相關特徵變數，並針對訓練詞嵌入模型的文本資料進行前處理。學習一個與電影相關的 Word2vec 詞嵌入模型為本研究的重點，透過電影評論文本的匯入訓練模型，並輸入意圖種子字透過詞嵌入模型進行擴充，預期生成意圖辭典提供使用者評論進行意圖分析。最後再將所有特徵變數透過機器學習的方法建構模型，進行最後的預測評估。

[11] https://www.imdb.com/?ref_=nv_home

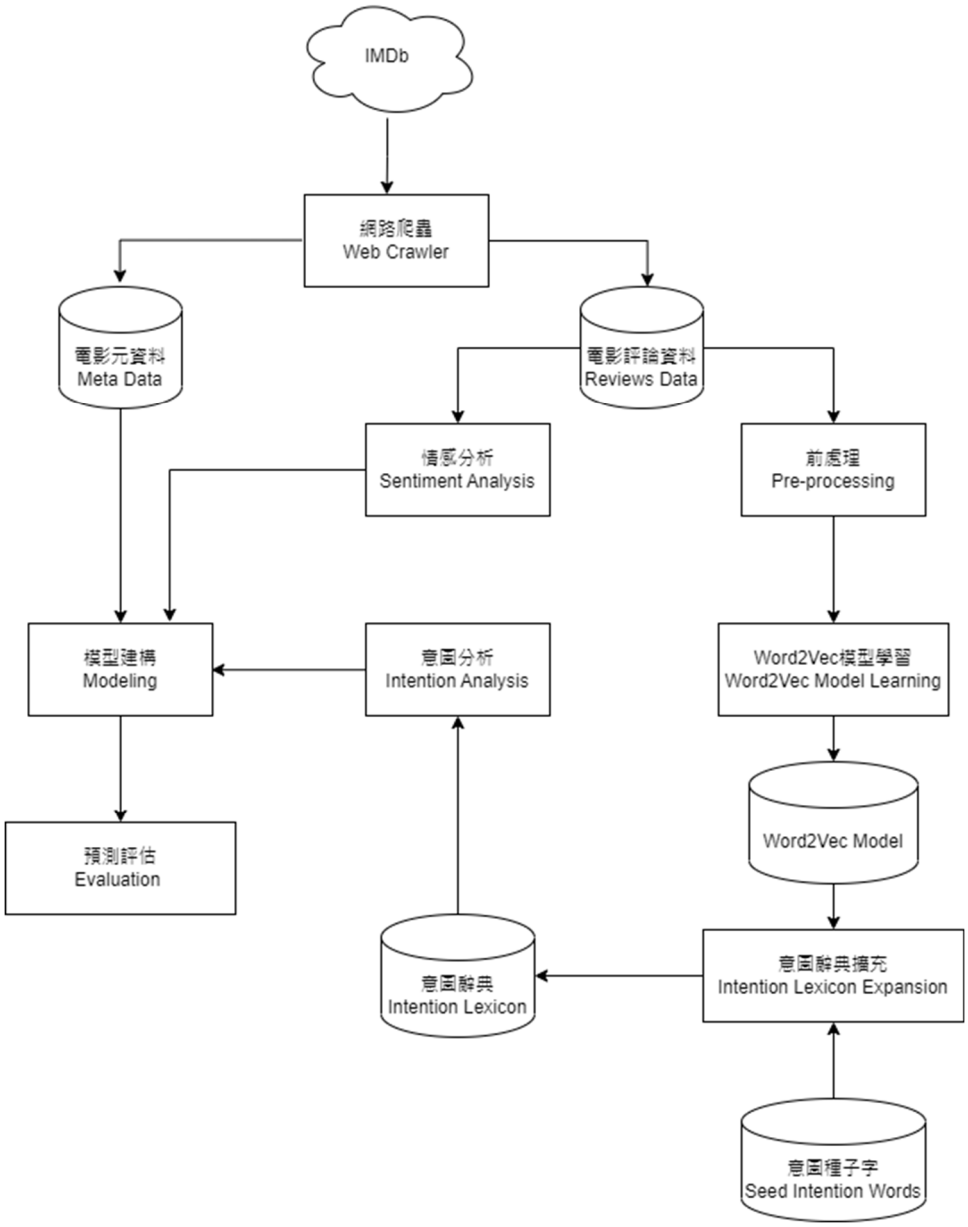

圖 1 研究流程圖

3.1 網路爬蟲

本研究所使用的數據集來自於亞馬遜公司(Amazon)旗下的 IMDb 網站，該網站是全球最大的電影資料庫網站，平台蒐集了多元的影視娛樂相關資訊，包含:電影、電視以及電子遊戲等。而電影在平台中有著最熱門的討論，也概括了豐富的電影相關資訊，如:演員、導演、劇情、投票評分、票房、影評等等訊息。IMDb 其中一項強大的功能是使用者可以針對影片進行評比，因為網站流量很大，因此會自動

過濾一些偏差值可能造成的影響，這也更好讓使用者在瀏覽電影時有著更加客觀的參考數據。

本研究爬蟲使用 Python 中 Beautiful Soup[12]以及 Selenium[13]剖析 HTML 標籤，進一步抓取網頁上靜態與動態資訊。資料抓取範圍包含電影元資料，也就是電影本身的相關資訊，以及每部電影的所有使用者評論。電影訓練資料集抓取範圍為 2020~2021 兩年的期間。考慮到後續資料研究的完整性，在過濾電影的時候設置了一些條件，包含:必須為長影片、至少超過 1000 人次以上投票、在美國上映過的電影，以及含有 Box Office 的四個值: 預算(Budget)、北美票房(Gross US & Canada)、北美首周周末票房(Opening weekend US & Canada)、全球票房(Gross worldwide)。至於缺失值的部分，是因為有些電影只在串流平台上架，因此無法計算票房損益。

這裡需要特別注意的是，由於進行 Word2vec[14]詞嵌入是需要大量文本進行訓練的，因此在使用者評論的數據爬取了兩個不同年份範圍的資料集，分別為與電影元資料相對應的評論資料集，用於最終的電影訓練資料。另一個資料爬取範圍為 2019~2022，為 Word2vec 詞嵌入模型訓練文本。因為只需要大量使用者評論文本作為模型訓練資料，並不考慮其評論內容來自哪部電影，因此並不需要額外設置條件來過濾。最後將爬蟲抓取下來的資料分別存成 CSV 格式，以便後續研究處理。

在 IMDb 網站上可以取得非常多元的電影相關屬性，本研究沿用了參考論文中提到相關的屬性，另外加上參考文章中未提及但有呈現在網站上的電影相關值，以下將介紹篩選過後用於本研究的變數。

- 電影分級(Certificate)

  IMDb 網頁呈現的電影分級綜合了許多不同國家的分級制度，為了研究的一致性，將採取分類與評級管理委員會(The Classification and Ratings Administration, CARA)所提出的一套分級制度作為基準，此委員會隸屬於美國電影協會(The Motion Picture Association of America, MPAA)，組織成立於 1922 年，主要成員由美國前七大影視傳媒巨頭共同組織而成。分級制度主要分為五種，分別是: G(普遍級)、 PG(輔導級，家長陪同觀看)、 PG-13(特別輔導級，建議 13 歲以

12 https://www.crummy.com/software/BeautifulSoup/bs4/doc/
13 https://selenium-python.readthedocs.io/
14 https://radimrehurek.com/gensim/models/word2vec.html

上兒童觀看)、 R(限制級，未滿 17 歲陪同觀看)、 NC-17(17 歲以下不可觀看)[15]。

- 電影續集(Sequel)

一部討論度高的作品，往往會因為受到觀眾的關注，亦或是票房亮眼的表現，讓製作方有計畫拍攝電影續集，以延伸原作品帶來的效益(吳坤宗 2010)。此變數由人工判別，若電影為某部作品的續作，則判斷此電影續集變數為 1，反之為 0。

- 電影時長(Runtime)

徐邦 (2015)曾經提到部分觀眾認為電影時長越長越好，這對消費者而言更能感到值回電影票價 。

- 電影類型(Genre)

不同的電影類型會吸引不同受眾的喜好，其中一部電影可能會同時具有一種以上的類型，從研究數據中一共整理出十七種電影類型，包含: 動作(Action)、冒險(Adventure)、喜劇(Comedy)、恐怖(Horror)、科幻(Sci-Fi)、犯罪(Crime)、劇情(Drama)、家庭(Family)、懸疑(Mystery)、奇幻(Fantasy)、人物傳記(Biography)、動畫(Animation)、浪漫(Romance)、驚悚(Thriller)、音樂(Music)、歷史(History)、體育(Sport)。

- 投票次數(Votes)

投票次數將紀錄共有多少人次為電影進行投票。

- IMDb Rating

使用者可以點擊電影頁進行投票，其值範圍落在 1~10 分。IMDb Rating 會考慮所有使用者對電影投下的分數再加總除以投票次數(Votes)，並以平均呈現。

- Metascore

Metascore 是來自一個 metacritic[16]的第三方網站所提供的數值，此平台會為每一部影視娛樂作品，包含:電視、電影、遊戲、音樂專輯，過濾至少四位評論

[15] https://en.wikipedia.org/wiki/Motion_Picture_Association_film_rating_system
[16] https://www.metacritic.com/

家，這些人是世界上有知名度且具有一定可信度的評論家，平台會運用加權平均值來總結他們的觀點。

- 製作國家(Countries of Origin)

此變數紀錄電影的製作國家，很多時候一部電影可能為多個國家團隊共同製作，研究數據中一共出現十七個不同的國家，包含: 美國(United States)、澳洲(Australia)、中國(China)、加拿大(Canada)、日本(Japan)、英國(United Kingdom)、南韓(South Korea)、德國(Germany)、南非(South Africa)、香港(Hong Kong)、盧森堡(Luxembourg)、法國(France)、瑞典(Sweden)、保加利亞(Bulgaria)、西班牙(Spain)、墨西哥(Mexico)、智利(Chile)。

- 使用者評論數(Reviews)

使用者評論數可以點擊電影專頁裡面查詢，同時也可以進一步查看每筆評論的內容。

以下四個變數為電影票房相關值，單位皆為美金。

- 預算(Budget)
- 北美票房(Gross US & Canada)
- 北美首周周末票房(Opening weekend US & Canada)
- 全球票房(Gross worldwide)

圖 2 與圖 3 為 IMDB 電影頁面，畫面中呈現許多電影相關資訊，研究透過網頁爬蟲抓取所有資訊，並挑選相關的變數作為電影元資料的數據集。圖 4 為使用者評論頁面，其內容可能包含使用者觀影後的心得，亦或是尚未前往觀影的期待及提問等等，一樣透過爬蟲的方式抓取使用者評論，並依照使用目的不同分為兩個數據集，分別是:與電影元資料匹配的使用者評論數據集，以及訓練 Word2vec 模型的使用者評論數據集。這些文字訊息是本研究很重要的實驗數據，預期能透過評論歸納出意圖內容與票房之間的影響。

圖 2~圖 4 紅色框分別代表:

1: 電影名稱、上映年分、電影分級、電影時長； 2: 電影類型； 3: 投票次數、IMDb Rating；

4: Metascore、使用者評論數； 5: 製作國家；

6: 預算、北美票房、北美首周周末票房、全球票房； 7: 使用者評論內容

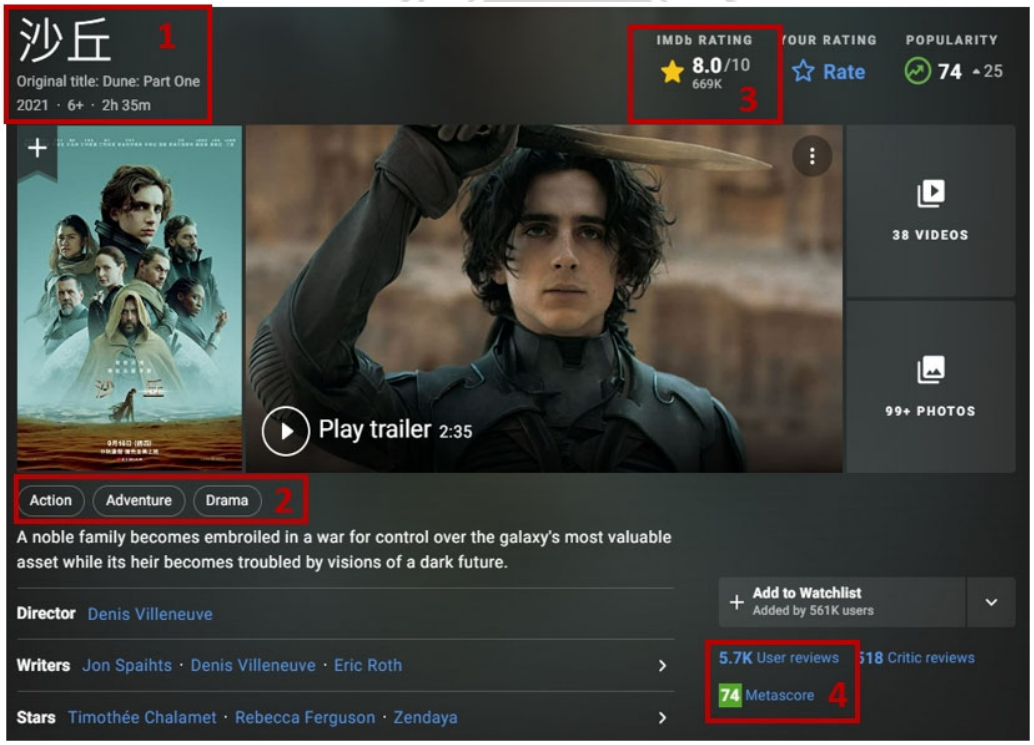

圖 2 IMDB 電影頁面示意圖[17]

[17] https://www.imdb.com/title/tt1160419/?ref_=rvi_tt

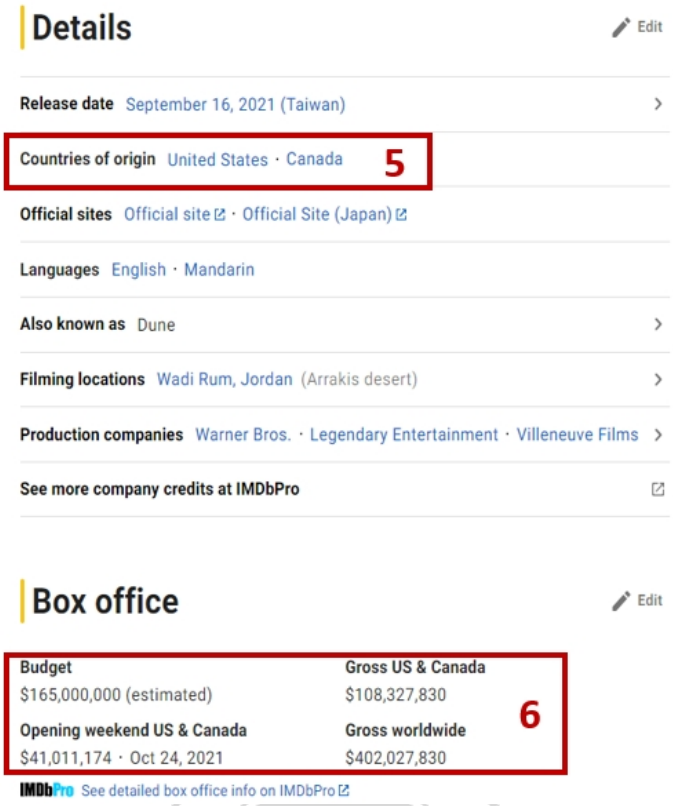

圖 3 IMDB 電影頁面示意圖

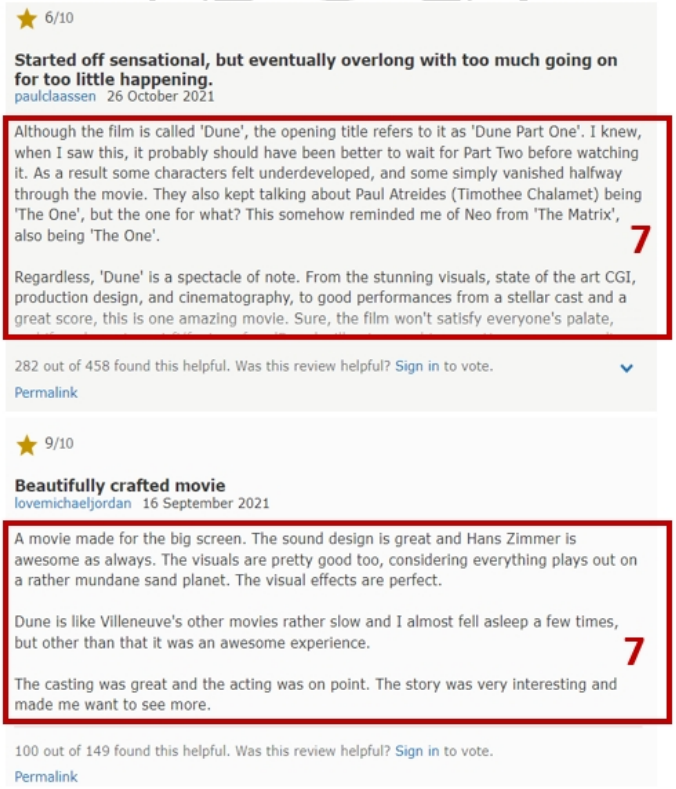

圖 4 IMDB 使用者評論頁面示意圖

3.2 資料前處理

針對 2019~2022 年間抓取的使用者評論進行資料前處理，所生成的文本作為下一階段 Word2vec 模型訓練的輸入。

- 非相關評論過濾

使用者評論中可能同時包含連結、表情符號或是一些雜訊等非文字的內容，因此在進行分析前必須先進行初步的過濾也確保文本皆為純文字。

- 大小寫轉換

因為使用者評論為英文文本，為了不讓單字有大小寫的區別，因此將把所有的字母全部轉換為小寫，讓不同大小寫卻一樣拼法的單字，都呈現相同意思。

- 斷句、斷詞

由於接下來要訓練的 Word2vec 模型文本必須以詞為單位，因此需要先將使用者評論拆解成數個句子，接下來再將句子斷成單詞，讓文本內容皆以詞為最小單位呈現。

- 詞形還原

在英文文本中，同一個單字的拼法會隨著不同的時態而有所改變。為了方便後續的處理，可以使用 Python 中的 Lemmatization 方法將單字的不同表示形式歸一化，以降低文本的複雜度並減少詞彙數量。這樣做可以使得文本更加清晰易懂，同時也有助於提高後續的文本分析的效率。

- 詞性標註(Part-of-Speech tagging, POS tagging)

當閱讀一段文本時，需要將每個單詞歸類到其所屬的詞性中，這就是詞性標註。詞性標註是自然語言處理 (Natural Language Processing, NLP) 中的一個重要任務，其目的是將文本中的每個單詞標記為其相應的詞性，例如: 名詞、動詞、形容詞、副詞、介詞、代詞等等。這些標記可用於分析文本的語法結構、進行文本分類、語言翻譯、信息擷取等 NLP 相關研究。

- Bigram 組合

Symeonidis, Peikos, and Arampatzis (2022)曾經提到，單詞數量的增加可以讓購買意圖更加明確，而動詞片語的組成更容易表現出使用者購買意圖。其中助動詞+動詞的動詞片語組合最能夠表達使用者的購買意圖，因為助動詞通常是用來輔助動詞的意思或者是指示動詞所表達的時態、情態、語氣等訊息。例如："should watch"中的"should"是助動詞，"watch"是動詞，組成了"should watch"這個動詞片語，表示"應該看"的意思。本研究將每個單詞依據其詞性遍歷，將符合上述規則的兩個單詞串接為一個新的詞組，形成 Bigram。若所遍及單字詞性不符合組成 Bigram 的條件，則將該單詞留下來作為 Onegram。最後，將所有經過處理的 Onegram 和 Bigram 以 CSV 檔案的形式存儲，以便作為 Word2vec 模型學習的輸入。

3.3 Word2vec 模型學習

Word2vec 是在 2013 年由 Google 的 Tomas Mikolov 等人提出的一套詞嵌入方法(Mikolov et al. 2013)，將字詞轉換為向量形式，並計算兩字詞之間的「餘弦值」，距離範圍為 0-1，值越大代表兩個詞關聯度越高。其主要可以分成連續詞袋模型(Continuous Bag Of Words, CBOW)以及 Skip-gram 模型，連續詞袋模型是透過已經既有的上下文，來預測字詞的機率，而 Skip-gram 則是透過字詞來預測上下文出現的機率，以上兩種模型都是計算詞向量的方法。也因為模型屬於非監督式學習，所以需要透過學習大量的文本，才能夠使結果更準確。研究使用 Python 中 Gensim 套件的 Word2Vec 模型進行訓練，並利用 Bigram 檔案匯入模型後，能夠在生成同義詞時涵蓋更多詞語的範圍，增加了詞彙的多樣性。

3.4 意圖辭典擴充

有了上一階段訓練好的 Word2vec 模型，接著將意圖種子字匯入到模型中進行迭帶，只考慮模型生成的前十個同義詞，再將每個種子字生成的前十個同義詞加上原本的種子字，刪除掉重複出現的詞後，作為第一次擴充的結果。第二次迭代匯入的種子字為第一次迭代的結果(只有 Iteration 1，不包含 Seed)，一樣將生成的前十個同義詞加上第一次擴充的結果(Seed + Iteration 1)，刪除掉重複出現的詞後，作為

第二次擴充的結果。依此類推可以重複擴充無限次，考量到詞彙數量多寡，本研究只進行兩次的迭帶。

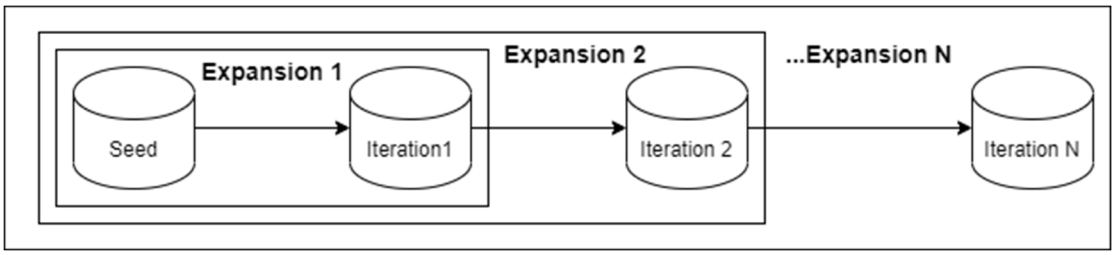

圖 5 意圖辭典擴充示意圖

3.5 意圖分析

本研究使用的分析方法是參照意圖探勘方法中的字典法，此階段使用的數據為與電影元資料相匹配的使用者評論資料，並對照意圖辭典中的意圖字。如果一筆評論裡出現一個或是多個符合辭典的字，則將該筆評論標註為 1，若皆無出現意圖字，則該筆評論標註為 0。接著，計算每部電影使用者評論中具有意圖的筆數，並生成對應的意圖特徵值。

3.6 情感分析

使用者評論中會間接表達自身對於事物的看法以及喜好程度，因此進一步針對文字進行情感分析。此階段情感分析的數據集為 2020~2021 年間與電影元資料相匹配的使用者評論資料集，使用 VARDER Sentiment (Hutto and Gilbert 2014)情感分析方法，情感值範圍落在-1 到 1 之間，分為三類 0 為中性，數值越接近 1 越正面，越接近-1 越負面。

考慮到熱門電影通常擁有更多的使用者評論，因此將根據評論數量的多寡給予不同的權重，以提高熱門電影使用者評論對分析的影響力。使用統計分位數的方法將使用者評論總數分為十個分位數，並選取第三和第七分位數作為劃分的標準。如果一部電影的評論數量小於三分位數，則給予權重 1；如果評論數量大於或等於三分位數，但小於七分位數，則給予權重 2；如果評論數量大於或等於七分位數，

則給予權重 4。這樣可以更有效地考慮電影討論度對使用者評論的影響。以下，式一將參考(Ahmad, Bakar, and Yaakub 2020)所提出計算情感分析權重的公式，其中 positive sentiment 為屬於正面情感的評論數量，negative sentiment 為負面情感的評論數量，w 為給予的權重值。

$$WPNratio = \frac{\text{positive sentiment}}{negative\ sentiment} * w \qquad \text{(式一)}$$

表 1 研究使用變數

電影分級(Certificate)	電影續集(Sequel)
電影時長(Runtime)	電影類型(Genre)
投票次數(Votes)	IMDB Rating
Metascore	製作國家(Countries of Origin)
使用者評論數(Reviews)	預算(Budget)
北美票房(Gross US & Canada)	全球票房(Gross worldwide)
北美首周周末票房(Opening weekend US & Canada)	
情感分析	意圖分析

3.7 模型建構

本研究最終目的是探討具有意圖的電影評論與電影票房之間的相關性，研究中所有變數皆為連續值或經由轉換過後的虛擬變數 ，因此皆採用回歸模型進行學習，以下將介紹四種機器學習分類法。

- 支援向量回歸(Support Vector Regression, SVR) (Drucker et al. 1996)

  支援向量回歸為支援向量機(Support Vector Machine, SVM)的延伸型態，能夠解決連續型的預測問題，此外分類器也提供了三種不同的核函數，其中包含線性與非線性模型，透過尋找最佳超平面將不同的資料分開，而最佳超平面即為其距離兩個類別的邊界達到最大，尋最大邊界值。

➢ 線性回歸(Linear Regression, LR) (Galton 1886)

線性回歸模型相對較簡單，易於透過數學公式的解釋來產生預測，同時也是一種資料分析技術，可以使用相關的已知資料來預測未知的值，線性回歸很適合應用於資料科學方法研究並解決複雜問題。

➢ 決策樹(Decision Tree, DT) (Quinlan 1992)

本研究採用回歸型的決策樹。決策樹為樹狀結構的模型，每個樹的節點皆為一個特徵，其節點可以再向下進行分支，並依循樹狀結構找到最後葉節點的分類。

➢ 隨機森林(Random Forest, RF) (Breiman 2001)

可以將隨機森林視為多個決策樹的組合，為 Bagging 算法也屬於集成學習(Ensemble Learning)的一種，每棵決策樹之間相互獨立，得到每棵樹的分類結果後，再透過投票的方式決定最終預測。

3.8 預測評估

本研究將資料進行十折交叉驗證(10-fold cross-validation)，把資料樣本切分成十等分，並依序取其中的一份資料做為測試集，其餘九份作為訓練集，一直重覆到每一份資料都做過測試集後停止，再將十次模型學習的數據取平均得到最終值。以下將介紹五種用於本研究回歸模型的評估指標。

➢ 皮爾森積動差相關係數(Pearson Correlation Coefficient, CC)

用於探討兩連續變數(X,Y)間的線性相關，值介於 1 到-1 之間。若兩變數間為正相關，則 X 越大，Y 越大;相反的，若變數間為負相關，則 X 越大，Y 越小。

$$r(x,y) = \frac{\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i-\overline{Y})^2}} \quad \text{(式二)}$$

$\overline{X}$為 X 平均數；$\overline{Y}$為 Y 平均數,n 為樣本數 0

- 平均絕對誤差(Mean Absolute Error, MAE)

每次測量的絕對誤差取決對值後求平均，可用來觀測預測值誤差的測量是否需要調整，其值越小模型預測能力越好，

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \qquad \text{(式三)}$$

y_i為預測值；$\hat{y}_i$為實際值

- 均方根誤差(Root Mean Square Error, RMSE)

均方根誤差也是變異數的平方根，主要用來聚集預測中誤差的大小，其值越小模型預測越好。

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \qquad \text{(式四)}$$

- 相對絕對誤差(Relative Absolute Error, RAE)

預測模型的定義為實際值和預測值之間的總誤差值，簡單模型為實際值與實際值的平均值之間總誤差值。相對絕對誤差就是檢查模型是否比簡單模型的表現還好，其值越小模型預測越好。

$$RAE = \frac{\sum_{i=1}^{n}|y_i - \hat{y}_i|}{\sum_{i=1}^{n}|y_i - \overline{y}|} \qquad \text{(式五)}$$

$\overline{y}$為真實值平均 0

- 相對平方根誤差(Root Relative Squared Error, RRSE)

模型預測值與真實值的平均值進行比較，來評估模型的表現好壞，其值越小模型預測越好。

$$RRSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}} \qquad \text{(式六)}$$

第四章 實驗結果與評估

本章節將詳細介紹研究的實驗結果。包含研究數據集的過濾條件、對應的年份範圍、數據量。接著介紹意圖種子字的過濾條件，以及本研究最終決定的意圖種子辭典。進行敘述統計，分別針對類別型以及連續型的電影相關變數做解析。最後針對實驗的數據集進行實驗結果分析，並調整了相關的變數來進行敏感度分析。

4.1 資料集

本研究數據集來源為 IMDb (Internet Movie Database)網路電影資料庫。參照表 2，研究選取時間段 2020~2021 年間的電影，利用 Python 中 selenium 動態網頁爬蟲的方式抓取所有電影相關資訊。一開始在 IMDb 網頁搜尋(Advanced Title Search)中先進行以下限制: 長影片(Feature Film)、至少超過 1000 人次以上投票(Number of Votes)、在美國上映過的電影(Countries)，在這個限制下網頁一共出現了 835 部電影,考慮到之後研究時變數的需求,在爬蟲抓取資料時只會保留包含四個 Box office 值的電影，四個值為: 預算(Budget)、北美票房(Gross US & Canada)、北美首周周末票房(Opening weekend US & Canada)、全球票房(Gross worldwide)，同時使用者評論(User Reviews)必須大於 100 筆，在過濾後一共蒐集了 91 部電影。使用者評論的部分分為兩種資料集，第一種是與電影元資料相匹配的使用者評論資料，用於使用意圖分析以及情感分析，針對過濾後的 91 部電影抓取各自的使用者評論，一共得到 130,985 筆；另一個使用者評論集用於 Word2vec 詞嵌入模型學習的文本，參照電影元資料在網頁中的搜尋限制，抓取 2019~2022 所有 IMDb 上的電影，因為只需要大量文本資料做訓練，因此不會考慮是否含有票房缺失值，並扣掉不包含使用者評論的電影，最後一共從 1,454 部電影中，得到了 618,125 筆評論。

表 2 研究資料集

資料集	資料筆數	說明
電影元資料集	91 筆電影資料	所有與電影相關的屬性資料，為實驗模型主要的數據集
使用者評論集 2020~2021	130,985 筆使用者評論	從電影元資料 91 部電影中抓取的使用者評論，用於情感分析、意圖分析
使用者評論集 2019~2022	618,125 筆使用者評論	為 Word2vec 模型的訓練文本

4.2 意圖種子字

文獻回顧中提到有些研究曾經提出與意圖相關的詞，整理後得到 28 個意圖詞: recommended, next time, again, looking, get, buy, want, need, purchase, use, replace, choose, take, find, like, thinking, would like, should buy, would choose, could get, could find, would use, would have, must buy, cannot wait, looking forward, keep an eye on, must have(Symeonidis, Peikos, and Arampatzis 2022; Ahmad et al. 2020; Zhang et al. 2021)，本研究將參考以上單詞，並過濾掉非電影意圖相關的字，最後留下 4 個字詞: recommend, next time, cannot wait, looking forward。參照以上意圖詞，另外在研究中加入 4 個新的字詞: must watch, would watch, should watch, watch again，以上 8 個字詞將作為本研究的電影意圖種子字，表 3 呈現本研究所使用的意圖種子字。

表 3 意圖種子字

recommend	next time
cannot wait	looking forward
must watch	would watch
should watch	watch again

4.3 敘述統計

本研究模型最終考慮了 15 個電影相關變數，表 4 與表 5 將變數分為離散型以及連續型，以下將詳細說明各變數的敘述統計結果。

表 4 呈現本研究中的 4 個離散變數。電影分級，依照 CARA 電影分級制度一共分為五個級別，分級為 PG-13 (特別輔導級，建議 13 歲以上兒童觀看)的電影最多，占總數的 44%，其次為 R (限制級，未滿 17 歲陪同觀看)占 40%、PG (輔導級，家長陪同觀看)占 14%，G (普遍級)只占 2%，另外在這 91 部電影中沒有任何一部電影屬於 NC-17 (17 歲以下不可觀看)的分級。電影續集的變數中有 54%的電影屬非續集的電影，有 46%的電影則以某部電影續集的形式發行，其中非續集電影的數量略高於續集電影。電影類型變數記錄著每一部電影的所屬類型，很多時候一部電影可能會包含多種電影類型，表格中可以看到有 73%的電影類型都屬於動作(Action)，其他超過 20%占比的電影類型依序是: 冒險(Adventure)、劇情(Drama)、喜劇(Comedy)、恐怖(Horror)，另外有五種電影類型占比不足 10%，包含:科幻(Sci-Fi)、傳記(Biography)、浪漫(Romance)、音樂(Music)、家庭(Family)。製作國家變數紀錄電影由哪些國家製作，而很多時候一部電影會由多國家的工作團隊共同監製，從表中可以得到結果，美國(United States)占比最多為 87%，其他只有 5 個國家占比超過 5%，分別為: 加拿大(Canada)、英國(United Kingdom)、中國(China)、澳洲(Australia)、日本(Japan)。

表 4 敘述統計表-離散變數

變數名稱	各類別占比
電影分級	G: 2(2%)；PG: 13(14%)；PG-13: 40(44%)； R: 36(40%)；NC-17: 0(0%)
電影續集	非續集: 49(54%)；續集: 42(46%)
電影類型	Action: 39(73%)；Adventure: 35(38%)；Drama: 32(35%)； Comedy: 29(32%)；Horror: 21(23%)；Crime: 16(18%)； Thriller: 14(15%)；Mystery: 13(14%)；Fantasy: 12(13%)； Animation: 10(11%)；Sci-Fi: 16(9%)；Biography: 7(8%)； Romance: 5(5%)；Music: 5(5%)；Family: 3(3%)； History: 2(2%)；Sport: 1(1%)
製作國家	United States: 79(87%)；Canada: 13(14%)； United Kingdom: 10(11%)；China: 8(9%)； Australia: 5(5%)；Japan: 5(5%)；Germany: 4(4%)； South Korea: 2(2%)；France: 2(2%)；Mexico: 2(2%)； South Africa: 1(1%)；Hong Kong: 1(1%)； Luxembourg: 1(1%)；Sweden: 1(1%)；Bulgaria: 1(1%)； Spain: 1(1%)；Chile: 1(1%)

表 5 為本研究中所有連續變數的敘述性統計。可以看到全球票房的值落差非常大，最小值為 266,963；最大值為 1,916,307,000；平均值為 151,174,900；標準差為 244,616,400，因為預測目標值的離散程度非常大，因此實驗的執行上有一定的難度。意圖分析在只使用種子字(Seed)的情況下，平均值為 87.18，這代表平均一部電影中含有意圖的評論約為 87 則，在經過一次擴充(Expansion 1)後，意圖評論次數的平均會成長到 604 則，第二次擴充(Expansion 2)會成長到 1,108 則。

表 5 敘述統計表-連續變數

變數名稱	平均值	標準差	最小值	最大值
電影時長(分鐘)	115.81	24.89	83	242
投票次數	127,967.74	138,270.93	5,706	749,188
IMDb Rating	6.27	0.90	3.90	8.30
Metascore	55.75	15.73	22	91
使用者評論數	1,439.40	1,585.04	121	8,070
預算(美元)	69,424,180	65,396,460	1,100,000	250,000,000
北美票房(美元)	59,060,400	96,515,830	93,147	814,115,100
北美首週週末票房(美元)	19,523,060	32,623,820	42,165	260,138,600
全球票房(美元)	151,174,900	244,616,400	266,963	1,916,307,000
情感分析	7.22	6.89	0.78	35.88
意圖分析(Seed)	87.18	93.17	4	480
意圖分析(Expansion 1)	604.88	707.74	41	3,495
意圖分析(Expansion 2)	1,108.81	1,239.02	93	5,951

表 6 為連續變數與全球票房的相關分析。從表中可以看到前兩名為北美票房、北美首周周末票房，這兩項的相關係數皆大於 0.9，說明各種票房類型之間有直接的高度相關。接著高於 0.5 的變數分別有:投票次數、意圖分析、預算以及使用者評論次數。

表 6 變數相關分析

變數名稱	相關係數
北美票房	0.9562
北美首周周末票房	0.9206
投票次數	0.7206
意圖分析(Seed)	0.6329
預算	0.6244
意圖分析(Expansion 2)	0.5252
意圖分析(Expansion 1)	0.5224
使用者評論次數	0.5085
情感分析	0.4194
IMDb Rating	0.3351
電影時長	0.2880
Metascore	0.1327

4.4 實驗結果

本研究以全球票房做為目標變數，本章節旨在詳細介紹各種 Word2vec 詞嵌入模型，所產生的意圖分析，於不同機器學習下的評估結果，其中包括四種演算法: 支援向量回歸(Support Vector Regression, SVR)、線性回歸(Linear Regression, LR)、決策樹(Decision Tree, DT)和隨機森林(Random Forest, RF)，以及五種模型評估指標: 皮爾森積動差相關係數(Pearson Correlation Coefficient, CC)、平均絕對誤差(Mean Absolute Error, MAE)、均方根誤差(Root Mean Square Error, RMSE)、相對絕對誤差(Relative Absolute Error, RAE)、相對平方根誤差(Root Relative Squared Error, RRSE)。

透過本章節的研究，將能夠深入瞭解不同機器學習模型的評估結果，以及不同 Word2vec 詞嵌入模型的性能表現。同時，還將探討意圖辭典對模型的影響，並評估刪除離群值對訓練的效果。這些結果將有助於本研究選擇最佳的模型和詞嵌入方法，以提高模型的準確性和可靠性。

4.4.1 原始數值

本小節使用 4.1 資料集挑選的 91 部電影作為研究數據。表 7 至表 10 顯示了四個不同的變數模型，分別為沒有加入意圖分析(No Intention)、只加入種子字的意圖分析(Seed)、只擴充一次意圖辭典的分析(Expansion 1)、擴充兩次意圖辭典的分析(Expansion 2)。根據表格結果，可以評估不同機器學習方法及其所使用的變數對於模型性能的影響。其中 SVR 模型在 CC、MAE、RAE 三個指標中，加入沒有意圖分析的"No Intention"變數時，獲得了最佳的表現。而在加入"Seed"變數的情況下，SVR 模型在 RMSE 和 RRSE 指標上達到最佳值。此外，在 LR 模型中，觀察到"No Intention"和"Seed"變數相對於其他 Expansion 結果，獲得了最佳的表現。在 DT 模型中，"No Intention"和"Expansion 2"兩個變數的指標結果皆達到相同的最佳結果。最後，在 RF 模型中，"Seed"和"Expansion 1"的變數在評估結果上呈現最佳性能。

綜合分析表 7 至表 10 的結果，可以得出，在這四個機器學習方法中，SVR 模型是最好的。特別是在 SVR 模型中，沒有加入意圖分析的"No Intention"變數呈現最佳表現，其次是加入"Seed"變數。如果只比較 Expansion 1 以及 Expansion 2 變數，使用 LR 與 DT 的方法都是 Expansion 2 結果較佳，在 RF 中則是相反，SVR 模型中兩個變數並沒有明顯任何一個較好。

表 7 支援向量回歸(SVR)模型評估結果

	No Intention	Seed	Expansion 1	Expansion 2
CC	**0.7217**	0.7169	0.7148	0.7156
MAE	**84,149,325**	86,710,802	84,611,610	85,337,776
RMSE	174,446,451	**173,571,081**	175,902,469	175,499,209
RAE	**59.0808**	60.8792	59.4054	59.9152
RRSE	70.8651	**70.5095**	71.4566	71.2928

註:同一個表格中，若模型的指標結果最佳，則以粗體字表示

表 8 線性回歸(LR)模型評估結果

	No Intention	Seed	Expansion 1	Expansion 2
CC	**0.6227**	0.6011	0.5841	0.6186
MAE	111,728,594	**111,212,172**	118,049,460	112,517,084
RMSE	**198,442,001**	200,904,771	204,910,461	199,151,292
RAE	78.4441	**78.0815**	82.8819	78.9977
RRSE	**80.6128**	81.6133	83.2405	80.9009

表 9 決策樹(DT)模型評估結果

	No Intention	Seed	Expansion 1	Expansion 2
CC	**0.6381**	0.582	0.5795	**0.6381**
MAE	**105,077,752**	105,895,437	107,354,936	**105,077,752**
RMSE	**192,458,687**	203,938,154	205,514,064	**192,458,687**
RAE	**73.7746**	74.3487	75.3734	**73.7746**
RRSE	**78.1822**	82.8455	83.4857	**78.1822**

表 10 隨機森林(RF)模型評估結果

	No Intention	Seed	Expansion 1	Expansion 2
CC	0.4855	0.5097	**0.5228**	0.5119
MAE	96,852,737	**94,225,344**	94,320,697	96,965,321
RMSE	213,996,624	209,607,933	**207,506,465**	209,232,317
RAE	67.9998	**66.1551**	66.2221	68.0789
RRSE	86.9315	85.1487	**84.2951**	84.9961

- Google, Wikipedia 預訓練模型

此處將使用不同 Word2vec 詞嵌入的預訓練模型進行比較，包括: Google 和 Wikipedia 的預訓練模型(Pre-trained Model) ，這些網站提供了豐富的新聞文章和百科全書條目，可用來擴充本研究的訓練數據。網路上已經存在許多預訓練的開源程式碼和通用模型可供使用，而不需要重新訓練整個模型。這樣的做法節省了時間和計算資源，同時也能受益於這些模型已經學習到的豐富語義和語境關係。

表 11 至表 14 呈現透過預訓練模型擴充意圖辭典，進而應用於意圖分析的變數，其中包含擴充一次以及兩次的模型(Google Expansion1 、Google Expansion 2、Wikipedia Expansion 1、Wikipedia Expansion 2)。根據表格數據，可以看出經由 Wikipedia 預訓練模型擴充出來的意圖分析，在 SVR 以及 LR 的學習方法上明顯地較 Google 預訓練模型還佳。Google 的預訓練模型使用 DT 學習有最好的效果，在 RF 的方法中 Google 和 Wikipedia 皆有良好的表現。透過觀察整個模型的評估結果表格，預訓練模型擴充一次或是兩次並沒有絕對最佳的結果，但從表 10 中可以得知 SVR 的數值皆優於其他三個機器學習的方法。

表 11 支援向量回歸(SVR)預訓練模型評估結果

	Google Expansion 1	**Google Expansion 2**	**Wikipedia Expansion 1**	**Wikipedia Expansion 2**
CC	0.7134	0.7087	**0.7213**	0.7097
MAE	86,365,802	86,359,442	**85,268,926**	86,317,795
RMSE	174,812,055	176,585,482	**173,913,106**	176,724,165
RAE	60.637	60.6325	**59.8669**	60.6033
RRSE	71.0137	71.7341	**70.6485**	71.7904

表 12 線性回歸(LR)預訓練模型評估結果

	Google Expansion 1	**Google Expansion 2**	**Wikipedia Expansion 1**	**Wikipedia Expansion 2**
CC	0.5876	0.5484	0.5941	**0.6272**
MAE	114,082,910	126,701,480	114,479,169	**109,857,002**
RMSE	203,462,909	219,643,199	203,062,614	**196,026,896**
RAE	80.097	88.9565	80.3752	**77.13**
RRSE	82.6523	89.2253	82.4898	**79.6317**

表 13 決策樹(DT)預訓練模型評估結果

	Google Expansion 1	Google Expansion 2	Wikipedia Expansion 1	Wikipedia Expansion 2
CC	0.5907	**0.6349**	0.5606	0.5697
MAE	**103,265,213**	106,706,762	109,595,549	110,905,246
RMSE	201,679,545	**194,960,560**	210,493,698	208,591,333
RAE	**72.502**	74.9183	76.9465	77.866
RRSE	81.928	**79.1986**	85.5086	84.7358

表 14 隨機森林(RF)預訓練模型評估結果

	Google Expansion 1	Google Expansion 2	Wikipedia Expansion 1	Wikipedia Expansion 2
CC	0.491	0.5254	**0.5305**	0.4899
MAE	97,817,148	**91,711,848**	92,989,574	97,465,475
RMSE	213,024,753	207,092,221	**206,324,883**	212,747,065
RAE	68.6769	**64.3904**	65.2875	68.43
RRSE	86.5367	84.126889	**83.8151**	86.4239

- Phrases 及 Unigram 模型訓練

在 Gensim 套件中，將利用 gensim.models.phrases[18]模塊進行詞嵌入模型的訓練。該模塊在訓練過程中會考慮更多的短語句，從而增加生成同義詞的多樣性和詞語涵蓋範圍。本研究將使用 Phrases 模塊建立一個新的 Word2vec 模型，利用意圖種子字，透過模型進行兩次意圖辭典的擴充，接著進一步觀察評估指標的結果並進行比較。另一方面因為本研究前面提及到的 Word2vec 詞嵌入模型都是使用 Bigram 文本作訓練，因此這部分想要訓練一個 Unigram 的 Word2vec 模型，進一步分析比較。

表 15 至 18 呈現四個不同模型擴充的結果，分別為 Phrases 以及 Unigram 的模型擴充一次和兩次的評估結果(Phrases Expansion 1、Phrases Expansion 2、

[18] https://radimrehurek.com/gensim/models/phrases.html

Unigram Expansion 1、Unigram Expansion 2)。可以觀察到，Unigram 的模型普遍優於 Phrases，Unigram 擴充兩次在 DT 以及 RF 的結果都是最佳的，LR 則是擴充一次最佳，而在 SVR 中也是偏向 Unigram 的結果較好。

表 15 支援向量回歸(SVR) Phrases 及 Unigram 模型評估結果

	Phrases Expansion 1	**Phrases Expansion 2**	**Unigram Expansion1**	**Unigram Expansion2**
CC	0.7175	0.7126	**0.7193**	0.7099
MAE	84,655,792	85,341,026	**84,467,152**	85,506,478
RMSE	175,245,229	176,106,538	**174,867,763**	176,391,129
RAE	**59.4364**	59.9175	59.4654	60.0337
RRSE	71.1896	71.5395	**71.0363**	71.6551

表 16 線性回歸(LR) Phrases 及 Unigram 模型評估結果

	Phrases Expansion 1	**Phrases Expansion 2**	**Unigram Expansion1**	**Unigram Expansion2**
CC	0.5926	0.6213	**0.6339**	0.6228
MAE	118,511,350	112,176,879	**103,736,268**	109,128,820
RMSE	203,782,297	198,890,909	**189,524,756**	198,360,684
RAE	83.2062	78.7588	**72.8327**	76.6188
RRSE	82.7822	80.7952	**76.9904**	80.5798

表 17 決策樹(DT) Phrases 及 Unigram 模型評估結果

	Phrases Expansion 1	**Phrases Expansion 2**	**Unigram Expansion1**	**Unigram Expansion2**
CC	0.6213	0.574	0.5629	**0.6341**
MAE	108,767,098	107,415,152	114,141,746	**105,319,884**
RMSE	196,811,629	207,632,098	212,014,099	**194,166,205**
RAE	76.3648	75.4156	80.1383	**73.9446**
RRSE	79.9505	84.3461	86.1262	**78.8759**

表 18 隨機森林(RF) Phrases 及 Unigram 模型評估結果

	Phrases Expansion 1	**Phrases Expansion 2**	**Unigram Expansion1**	**Unigram Expansion2**
CC	0.4688	0.5147	0.5022	**0.5253**
MAE	98,338,062	92,524,459	96,473,867	**91,764,581**
RMSE	216,948,346	209,101,486	211,139,136	**207,135,197**
RAE	69.0427	64.961	67.7338	**64.4274**
RRSE	88.1306	84.943	85.7708	**84.1442**

表 19 至表 22 為以上表 7 到 18 的彙整表，從每個模型中挑選一個最佳的機器學習方法及其評估結果做為比較。從各表中可以知道在原始數值當中 SVR 於各種模型的評估效果都是最適合的。在 SVR 中在沒有意圖分析的模型(No Intention)以及意圖種子字模型(Seed)各有優劣；LR 的模型中 Unigram 擴充一次的模型(Unigram Expansion 1)有著最佳的效果；D 則是 No Intention 和本研究提出的模型方法為擴充兩次(Expansion 2)有相同最佳結果；在 RF 中預訓練模型(Google Expansion 2、Wikipedia Expansion 1)的數值較佳。整體而言，在原始數值中就數量來說，模型擴充兩次會多於只擴充一次。

表 19 支援向量回歸(SVR)模型評估結果彙整表

	No Intention	Seed	Expansion 2	Google 1	Wiki 1	Phrases 1	Unigram 1
CC	**0.7217**	0.7169	0.7156	0.7134	0.7213	0.7175	0.7193
MAE	**84,149,325**	86,710,802	85,337,776	86,365,802	85,268,926	84,655,792	84,467,152
RMSE	174,446,451	**173,571,081**	175,499,209	174,812,055	173,913,106	175,245,229	174,867,763
RAE	**59.0808**	60.8792	59.9152	60.637	59.8669	59.4364	59.4654
RRSE	70.8651	**70.5095**	71.2928	71.0137	70.6485	71.1896	71.0363

註: Google 1/2 為 Google Expansion 1/2；Wiki 1/2 為 Wikipedia Expansion 1/2；Phrases1/2 為 Phrases Expansion 1/2；Unigram 1/2 為 Unigram Expansion 1/2

表 20 線性回歸(LR)模型評估結果彙整表

	No Intention	Seed	Expansion 2	Google 1	Wiki 2	Phrases 2	Unigram 1
CC	0.6227	0.6011	0.6186	0.5876	0.6272	0.6213	**0.6339**
MAE	111,728,594	111,212,172	112,517,084	114,082,910	109,857,002	112,176,879	**103,736,268**
RMSE	198,442,001	200,904,771	199,151,292	203,462,909	196,026,896	198,890,909	**189,524,756**
RAE	78.4441	78.0815	78.9977	80.097	77.13	78.7588	**72.8327**
RRSE	80.6128	81.6133	80.9009	82.6523	79.6317	80.7952	**76.9904**

表 21 決策樹(DT)模型評估結果彙整表

	No Intention	Seed	Expansion 2	Google 2	Wiki 2	Phrases 1	Unigram 2
CC	**0.6381**	0.582	**0.6381**	0.6349	0.5697	0.6213	0.6341
MAE	**105,077,752**	105,895,437	**105,077,752**	106,706,762	110,905,246	108,767,098	105,319,884
RMSE	**192,458,687**	203,938,154	**192,458,687**	194,960,560	208,591,333	196,811,629	194,166,205
RAE	**73.7746**	74.3487	**73.7746**	74.9183	77.866	76.3648	73.9446
RRSE	**78.1822**	82.8455	**78.1822**	79.1986	84.7358	79.9505	78.8759

表 22 隨機森林(RF)模型評估結果彙整表

	No Intention	Seed	Expansion 1	Google 2	Wiki 1	Phrases 2	Unigram 2
CC	0.4855	0.5097	0.5228	0.5254	**0.5305**	0.5147	0.5253
MAE	96,852,737	94,225,344	94,320,697	**91,711,848**	92,989,574	92,524,459	91,764,581
RMSE	213,996,624	209,607,933	207,506,465	207,092,221	**206,324,883**	209,101,486	207,135,197
RAE	67.9998	66.1551	66.2221	**64.3904**	65.2875	64.961	64.4274
RRSE	86.9315	85.1487	84.2951	84.1268	**83.8151**	84.943	84.1442

4.4.2 過濾後數值

觀察表 5 敘述統計表-連續變數，發現在全球票房數據中存在一些數值差異很大的離群值，為了提升模型的訓練效果，因此嘗試刪除這些離群值。總共刪除全球票房小於 10,000,000 或大於 1,000,000,000 的 10 部電影作品，最終保留 81 部電影作為本小節的訓練資料樣本。以下將介紹過濾後的電影樣本進行機器學習的模型評估指標。

表 23 至 26 可以看到在 SVR 以及 LR 中意圖種子字(Seed)的模型效果最佳，在 DT 的方法中則是沒有加入意圖(No Intention)的結果最佳，RF 方法中並沒有哪一個模型有絕對較好的效果。如果單看擴充的模型可以發現，LR 與 DT 都是擴充兩次的(Expansion 2)效果較好，SVR 則是擴充一次(Expansion 1)較好。

表 23 支援向量回歸(SVR)模型評估結果(過濾後)

	No Intention	Seed	Expansion 1	Expansion 2
CC	0.7171	**0.7293**	0.7146	0.7128
MAE	73,199,210	**72,172,107**	74,440,388	74,881,059
RMSE	112,283,671	**109,790,796**	112,711,444	113,260,079
RAE	61.4745	**60.6119**	62.5169	62.887
RRSE	69.6634	**68.1168**	69.9288	70.2692

表 24 線性回歸(LR)模型評估結果(過濾後)

	No Intention	Seed	Expansion 1	Expansion 2
CC	**0.6805**	0.678	0.6611	0.6735
MAE	89,093,453	**81,627,450**	90,961,668	88,208,570
RMSE	125,940,820	**120,577,744**	130,681,832	125,860,702
RAE	74.8229	**68.5528**	76.3919	74.0797
RRSE	78.1367	**74.8093**	81.0781	78.087

表 25 決策樹(DT)模型評估結果(過濾後)

	No Intention	Seed	Expansion 1	Expansion 2
CC	0.6933	0.6758	0.6902	**0.6935**
MAE	**73,625,268**	73,965,512	73,643,838	73,887,358
RMSE	**116,950,429**	119,142,205	117,230,239	116,992,785
RAE	**61.8323**	62.1181	61.8479	62.0524
RRSE	**72.5588**	73.9186	72.7324	72.5851

表 26 隨機森林(RF)模型評估結果(過濾後)

	No Intention	Seed	Expansion 1	Expansion 2
CC	0.6857	**0.6897**	0.6798	0.6851
MAE	**74,971,525**	76,351,235	75,639,867	75,866,525
RMSE	116,811,006	**116,140,723**	117,463,139	116,765,799
RAE	**62.963**	64.1217	63.5242	63.7146
RRSE	72.4723	**72.0565**	72.8769	72.4443

- Google, Wikipedia 預訓練模型(過濾後)

觀察表 27 至 30，SVR 和 LR 的方法下使用 Google 預訓練模型並且擴充一次(Google Expansion 1)的效果最佳，DT 的方法下使用 Wikipedia 預訓練模型並擴充一次(Wikipedia Expansion 1)的效果最佳，在 RF 中 Google 和 Wikipedia 擴充兩次的效果都不錯，但沒有一個是絕對最好。

表 27 支援向量回歸(SVR)預訓練模型評估結果(過濾後)

	Google Expansion 1	Google Expansion 2	Wikipedia Expansion 1	Wikipedia Expansion 2
CC	**0.727**	0.708	0.7191	0.7062
MAE	**72,135,746**	74,633,530	73,191,600	75,541,615
RMSE	**110,206,391**	113,865,042	111,547,083	114,348,537
RAE	**60.5814**	62.6791	61.4681	63.4417
RRSE	**68.3747**	70.6446	69.2065	70.9445

表 28 線性回歸(LR)預訓練模型評估結果(過濾後)

	Google Expansion 1	Google Expansion 2	Wikipedia Expansion 1	Wikipedia Expansion 2
CC	0.6633	**0.6821**	0.6691	0.6785
MAE	**84,371,201**	88,340,479	87,898,420	88,384,580
RMSE	**123,669,321**	124,195,279	126,384,957	125,742,748
RAE	**70.857**	74.1905	73.8193	74.2276
RRSE	**76.7274**	77.0537	78.4122	78.0138

表 29 決策樹(DT)預訓練模型評估結果(過濾後)

	Google Expansion 1	Google Expansion 2	Wikipedia Expansion 1	Wikipedia Expansion 2
CC	0.6794	0.6852	**0.6882**	0.6812
MAE	74,023,924	75,119,603	**72,128,804**	73,873,879
RMSE	118,299,670	118,303,796	**117,227,957**	118,077,869
RAE	62.1671	63.0873	**60.5756**	62.0411
RRSE	73.3959	73.3985	**72.731**	73.2583

表 30 隨機森林(RF)預訓練模型評估結果(過濾後)

	Google Expansion 1	Google Expansion 2	Wikipedia Expansion 1	Wikipedia Expansion 2
CC	0.6483	**0.6916**	0.6709	0.6747
MAE	77,861,479	75,009,110	76,294,913	**74,841,969**
RMSE	121,865,353	**115,920,267**	118,761,557	118,134,099
RAE	65.39	62.9945	64.0744	**62.8541**
RRSE	75.6081	**71.9197**	73.6825	73.2932

- Phrases 及 Unigram 模型訓練

從表 31 至 34 觀察到， SVR 的數值在所有模型中為效果最佳的方法。在 SVR 中 Phrases 擴充一次(Phrases Expansion 1)的結果最好，在 DT 中是 Unigram 擴充兩次(Unigram Expansion 2)結果最好，在 LR 及 RF 則是 Phrases 及 Unigram 各有優劣。

表 31 支援向量回歸(SVR) Phrases 及 Unigram 模型評估結果(過濾後)

	Phrases Expansion 1	**Phrases Expansion 2**	**Unigram Expansion1**	**Unigram Expansion2**
CC	**0.7147**	0.7142	0.7142	0.7083
MAE	**74,496,431**	74,942,934	74,918,448	75,406,500
RMSE	**112,597,441**	113,020,834	112,811,575	113,903,526
RAE	**62.564**	62.9389	62.9184	63.3283
RRSE	**69.8581**	70.1208	69.9910	70.6684

表 32 線性回歸(LR) Phrases 及 Unigram 模型評估結果(過濾後)

	Phrases Expansion 1	**Phrases Expansion 2**	**Unigram Expansion1**	**Unigram Expansion2**
CC	0.6847	0.698	**0.7015**	0.6469
MAE	88,897,937	**83,629,490**	85,408,431	91,920,810
RMSE	125,401,957	120,367,669	**118,956,115**	132,143,596
RAE	74.6587	**70.2341**	71.7281	77.1974
RRSE	77.8023	74.679	**73.8032**	81.9850

表 33 決策樹(DT) Phrases 及 Unigram 模型評估結果(過濾後)

	Phrases Expansion 1	**Phrases Expansion 2**	**Unigram Expansion1**	**Unigram Expansion2**
CC	0.6773	0.6956	0.6882	**0.7016**
MAE	75,804,818	73,017,590	75,652,843	**72,666,941**
RMSE	119,069,253	116,379,611	118,062,763	**115,170,732**
RAE	63.6628	61.322	63.5351	**61.0275**
RRSE	73.8734	72.2047	73.2489	**71.4546**

表 34 隨機森林(RF) Phrases 及 Unigram 模型評估結果(過濾後)

	Phrases Expansion 1	**Phrases Expansion 2**	**Unigram Expansion1**	**Unigram Expansion2**
CC	0.6897	0.6791	0.6769	**0.6958**
MAE	**72,165,674**	76,329,634	76,554,258	72,860,330
RMSE	116,128,313	117,541,541	117,945,426	**115,386,996**
RAE	**60.6065**	64.1035	64.2922	61.1899
RRSE	72.0488	72.9256	73.1761	**71.5888**

表 35 至 38 為以上表 23 到 34 的彙整表，從每個模型中挑選一個最佳的機器學習方法及其評估指標結果做比較。從以下結果中可以知道 SVR 的指標結果為所有方法中最佳。在 SVR 中意圖種子字模型(Seed)以及 Google 擴充一次(Google Expansion 1)的結果較好；在 LR 中是 Seed 和 Unigram 擴充一次(Unigram Expansion 1)結果較好；DT 的方法中是 Wikipedia 以及 Unigram 擴充兩次(Wikipedia Expansion 2、Unigram Expansion 2) 的效果最佳；RF 則是 Phrases 擴充一次(Phrases Expansion 1)以及 Unigram 擴充兩次的效果較佳。整體來看，過濾後數值的模型就數量而言擴充兩次的次數會多於只擴充一次。

表 35 支援向量回歸(SVR)模型評估結果彙整表(過濾後)

	No Intention	Seed	Expansion 1	Google 1	Wiki 1	Phrases 1	Unigram 1
CC	0.7171	**0.7293**	0.7146	0.727	0.7191	0.6847	0.7142
MAE	73,199,210	72,172,107	74,440,388	**72,135,746**	73,191,600	88,897,937	74,918,448
RMSE	112,283,671	**109,790,796**	112,711,444	110,206,391	111,547,083	125,401,957	112,811,575
RAE	61.4745	60.6119	62.5169	**60.5814**	61.4681	74.6587	62.9184
RRSE	69.6634	**68.1168**	69.9288	68.3747	69.2065	77.8023	69.9910

表 36 線性回歸(LR)模型評估結果彙整表(過濾後)

	No Intention	Seed	Expansion 2	Google 1	Wiki 2	Phrases 2	Unigram 1
CC	0.6805	0.678	0.6735	0.6633	0.6785	0.698	**0.7015**
MAE	89,093,453	**81,627,450**	88,208,570	84,371,201	88,384,580	83,629,490	85,408,431
RMSE	125,940,820	120,577,744	125,860,702	123,669,321	125,742,748	120,367,669	**118,956,115**
RAE	74.8229	**68.5528**	74.0797	70.857	74.2276	70.2341	71.7281
RRSE	78.1367	74.8093	78.087	76.7274	78.0138	74.679	**73.8032**

表 37 決策樹(DT)模型評估結果彙整表(過濾後)

	No Intention	Seed	Expansion 2	Google 1	Wiki 2	Phrases 2	Unigram 2
CC	0.6933	0.6758	0.6935	0.6794	0.6882	0.6956	**0.7016**
MAE	73,625,268	73,965,512	73,887,358	74,023,924	**72,128,804**	73,017,590	72,666,941
RMSE	116,950,429	119,142,205	116,992,785	118,299,670	117,227,957	116,379,611	**115,170,732**
RAE	61.8323	62.1181	62.0524	62.1671	**60.5756**	61.322	61.0275
RRSE	72.5588	73.9186	72.5851	73.3959	72.731	72.2047	**71.4546**

表 38 隨機森林(RF)模型評估結果彙整表(過濾後)

	No Intention	Seed	Expansion 2	Google 2	Wiki 2	Phrases 1	Unigram 2
CC	0.6857	0.6897	0.6851	0.6916	0.6747	0.6897	**0.6958**
MAE	74,971,525	76,351,235	75,866,525	75,009,110	74,841,969	**72,165,674**	72,860,330
RMSE	116,811,006	116,140,723	116,765,799	115,920,267	118,134,099	116,128,313	**115,386,996**
RAE	62.963	64.1217	63.7146	62.9945	62.8541	**60.6065**	61.1899
RRSE	72.4723	72.0565	72.4443	71.9197	73.2932	72.0488	**71.5888**

綜合以上原始數值和過濾後數值可以觀察到，在這兩種資料集當中支援向量回歸(SVR)的模型評估結果都是最佳的，並且在同一種模型當中，以數量而言擴充兩次的會多於只擴充一次的。

4.5 敏感度分析

本章節使用原始數值數據集，討論關於意圖種子字的數量、北美票房以及北美首周票房的敏感度分析。

4.5.1 意圖種子字數量

參照 4.2 意圖種子字，可以得知本研究論文使用的八個電影意圖種子字是由原先三篇參考論文提出的意圖種子字過濾而來。本小節針對原先的二十八個意圖種子字，直接進行意圖分析，並進一步觀察結果。

表 39 至 42 為使用二十八個意圖種子字進行的意圖模型學習結果。可以看到 SVR 的數值為四種方法中最佳，在 SVR 中不使用意圖種子字的模型(No Intention)效果最佳；LR 以及 RF 最佳結果皆為 Unigram 擴充一次(Unigram Expansion 1)；DT 則是 Wikipedia 擴充一次(Wikipedia Expansion 1)的結果最好。整體而言，在二十八個意圖字的模型當中，以數量來看擴充一次多於擴充兩次。

表 39 支援向量回歸(SVR)模型評估結果彙整表(二十八個意圖字)

	No Intention	Seed	Expansion 2	Google 2	Wiki 1	Phrases 2	Unigram 1
CC	**0.7217**	0.7148	0.7144	0.7174	0.7166	0.7160	0.7179
MAE	**84,149,325**	85,947,716	85,484,580	85,067,596	84,920,588	85,168,153	85,109,875
RMSE	**174,446,451**	175,402,190	175,666,743	175,154,621	175,187,262	116,128,313	174,943,421
RAE	**59.0808**	60.3435	60.0183	59.7255	59.6223	59.7961	59.7552
RRSE	**70.8651**	71.2534	71.3609	71.1528	71.1661	71.2862	71.067

表 40 線性回歸(LR)模型評估結果彙整表(二十八個意圖字)

	No Intention	Seed	Expansion 1	Google 1	Wiki 2	Phrases 1	Unigram 1
CC	0.6227	0.6369	0.6319	0.6192	0.7166	0.6313	**0.6312**
MAE	111,728,594	108,560,134	107,280,387	110,757,793	113,523,942	107,626,139	**106,748,384**
RMSE	198,442,001	193,235,600	195,129,308	199,300,502	199,347,078	195,153,040	**194,357,626**
RAE	78.4441	76.2195	75.321	77.7625	79.7046	75.5638	**74.9475**
RRSE	80.6128	78.4978	79.2671	80.9616	80.9805	79.2767	**78.9536**

表 41 決策樹(DT)模型評估結果彙整表(二十八個意圖字)

	No Intention	Seed	Expansion 1	Google 1	Wiki 1	Phrases 1	Unigram 1
CC	0.6381	0.6437	0.6408	0.6433	**0.6450**	0.6380	0.6438
MAE	105,077,752	104,661,245	104,737,651	103,938,448	**103,475,098**	106,045,280	104,040,431
RMSE	192,458,687	190,481,873	191,478,062	190,213,395	**189,596,839**	191,974,230	190,181,074
RAE	73.7746	73.4821	73.5358	72.9747	**72.6493**	74.4539	73.0463
RRSE	78.1822	77.3792	77.7839	77.2701	**77.0197**	77.9854	77.2570

表 42 隨機森林(RF)模型評估結果彙整表(二十八個意圖字)

	No Intention	Seed	Expansion 2	Google 1	Wiki 1	Phrases 1	Unigram 1
CC	0.4855	0.4944	0.5195	0.5034	0.5046	0.5220	**0.5330**
MAE	96,852,737	95,801,116	93,836,488	94,252,521	94,915,204	95,826,005	**91,857,568**
RMSE	213,996,624	212,793,482	208,097,286	211,174,357	210,968,053	207,849,833	**205,894,457**
RAE	67.9998	67.2615	65.8821	66.1742	66.6395	67.279	**64.4927**
RRSE	86.9315	86.4428	84.5351	85.7851	85.7013	84.4345	**83.6402**

4.5.2 北美票房

本小節以北美票房替換全球票房，做為目標變數。觀察表 43 至 46，發現 SVR 為所有方法中數值最佳，並且在本研究提出的模型擴充一次後(Expansion 1)效果最佳；LR 中的 Wikipedia 擴充兩次(Wikipedia Expansion 2)結果最好；DT 是不加入任何意圖變數(No Intention)的值最好；RF 則是 Wikipedia 以及 Unigram 擴充一次(Wikipedia Expansion 1、Unigram Expansion 1)的結果較好。整體而言，在北美票房的變數當中，以數量來看模型擴充兩次會多於只擴充一次。

表 43 支援向量回歸(SVR)模型評估結果彙整表(北美票房)

	No Intention	Seed	Expansion 1	Google 2	Wiki 2	Phrases 1	Unigram 1
CC	0.5164	0.4938	**0.5706**	0.5139	0.5396	0.5665	0.5391
MAE	42,486,717	44,460,244	**40,259,614**	42,264,520	41,896,969	40,425,268	40,775,646
RMSE	82,489,611	84,083,149	**79,134,340**	82,650,665	81,021,383	79,7108	81,024,761
RAE	80.6226	84.3675	**76.3964**	80.2009	79.5035	76.7108	77.3757
RRSE	84.8321	86.4709	**81.3815**	84.9977	83.3222	81.5732	83.3256

表 44 線性回歸(LR)模型評估結果彙整表(北美票房)

	No Intention	Seed	Expansion 2	Google 2	Wiki 2	Phrases 2	Unigram 1
CC	0.534	0.5352	0.5626	0.5109	**0.5751**	0.5407	0.5621
MAE	50,751,931	46,156,884	45,865,400	46,937,876	**44,480,208**	49,269,982	44,209,923
RMSE	86,144,997	85,014,952	82,239,384	84,679,332	**79,556,070**	84,975,987	80,144,086
RAE	96.3066	87.5871	87.034	89.0691	**84.4054**	93.4945	83.8925
RRSE	88.5913	87.4291	84.5748	87.084	**81.8152**	87.3891	82.42

表 45 決策樹(DT)模型評估結果彙整表(北美票房)

	No Intention	Seed	Expansion 2	Google 1	Wiki 1	Phrases 2	Unigram 2
CC	**0.4996**	0.4799	0.4887	0.4898	0.4847	0.4765	0.4693
MAE	**45,087,034**	45,143,222	46,682,285	45,992,936	46,086,101	46,962,031	48,013,819
RMSE	**86,015,703**	87,041,844	86,789,860	86,394,930	87,738,573	88,211,228	88,534,640
RAE	**85.5569**	85.6635	88.5841	87.276	87.4528	89.1149	91.1108
RRSE	**88.4583**	89.5136	89.2545	88.8483	90.2301	90.7162	91.0488

表 46 隨機森林(RF)模型評估結果彙整表(北美票房)

	No Intention	Seed	Expansion 1	Google 2	Wiki 1	Phrases 2	Unigram 1
CC	0.4257	0.4136	0.3968	0.4144	**0.4389**	0.4152	0.4366
MAE	38,426,739	36,997,435	38,775,659	37,583,284	36,716,882	37,616,309	**36,538,197**
RMSE	87,308,541	87,830,747	88,939,719	87,884,217	**86,450,060**	87,814,959	86,503,271
RAE	72.9184	70.2061	73.5805	71.3178	69.6738	71.3805	**69.3347**
RRSE	89.7879	90.3249	91.4654	90.3799	**88.905**	90.3087	88.9597

4.5.3 北美首周周末票房

本小節以北美首周周末票房做為目標變數。觀察表 47 至 50，SVR 的結果為所有方法中最好的，其中 Phrases 以集 Unigram 在擴充一次(Phrases Expansion 1、Unigram Expansion 1) 時都有不錯的結果；LR 以及 DT 皆為預訓練模型(Wikipedia Expansion 1、Google Expansion 2)的效果較佳；而 RF 方法中則是不使用意圖的模型(No Intention)結果最好。整體而言在北美首周周末票房的變數中，以數量來看擴充兩次的模型會多於只擴充一次。

表 47 支援向量回歸(SVR)模型評估結果彙整表(北美首周周末票房)

	No Intention	Seed	Expansion 1	Google 2	Wiki 2	Phrases 1	Unigram 1
CC	0.611	0.5846	0.6273	0.5882	0.6166	**0.6378**	0.6283
MAE	15,033,651	15,532,974	14,889,710	15,498,112	15,004,708	14,772,440	**14,698,834**
RMSE	25,925,845	26,498,415	25,511,530	26,437,033	25,806,001	**25,285,855**	25,535,264
RAE	78.6924	81.3061	77.939	81.1236	78.5409	77.3251	**76.9398**
RRSE	78.8502	80.5916	77.5901	80.4049	78.4857	**76.9037**	77.6623

表 48 線性回歸(LR)模型評估結果彙整表(北美首周周末票房)

	No Intention	Seed	Expansion 2	Google 2	Wiki 1	Phrases 2	Unigram 2
CC	**0.6077**	0.544	0.5947	0.5876	0.5523	0.589	0.5406
MAE	16,407,707	16,913,331	16,940,752	16,046,358	**15,640,006**	16,992,644	17,887,245
RMSE	26,682,864	28,959,060	27,034,514	27,889,725	**26,258,315**	27,642,260	30,200,478
RAE	85.8848	88.5315	88,6750	83.9934	**81.8663**	88.9466	93.6293
RRSE	81.1526	88.0753	82,2221	84.8231	**79.8614**	84.0705	91.851

表 49 決策樹(DT)模型評估結果彙整表(北美首周周末票房)

	No Intention	Seed	Expansion 2	Google 2	Wiki 1	Phrases 2	Unigram 1
CC	0.5198	0.5924	0.559	**0.6065**	0.5714	0.5545	0.5576
MAE	15,605,522	14,975,100	15,399,591	**14,263,166**	14,690,946	14,555,463	14,856,864
RMSE	28,864,125	26,927,699	27,905,112	**26,025,331**	27,296,533	27,651,080	27,160,797
RAE	81.6858	78.3859	80.6079	**74.6594**	76.8986	76.1894	77.767
RRSE	87.7866	81.8972	84.8699	**79.1528**	83.019	84.0973	82.6061

表 50 隨機森林(RF)模型評估結果彙整表(北美首周周末票房)

	No Intention	Seed	Expansion 2	Google 2	Wiki 1	Phrases 1	Unigram 1
CC	**0.4706**	0.4308	0.4333	0.4506	0.4239	0.4285	0.4085
MAE	**13,806,495**	14,220,663	14,181,670	13,672,495	14,188,381	14,241,745	14,272,064
RMSE	**28,667,291**	29,539,970	29,412,806	29,040,562	29,727,615	29,674,514	29,969,918
RAE	**72.269**	74.4369	74.2328	71.5676	74.2679	74.5472	74.706
RRSE	**87.188**	89.8421	89.4553	88.3232	90.4128	90.2513	91.1497

第五章 結論與未來研究方向

5.1 結論

本研究抓取 IMDb 電影論壇平台上與電影相關的變數，並分析其與電影票房之間的關係，實驗更以意圖辭典的擴充為主要討論的方向，利用參考的文獻內容整理出相關的購買意圖種子字，並針對電影的領域知識(Domain Knowledge)將意圖種子字做了增減，形成了意圖種子辭典。本實驗提出自行訓練一個電影專屬的 Word2vec 詞嵌入模型，考慮到所讀文獻中曾提到，Bigram 的組合更容易表達使用者的購買意圖，因此在訓練文本中將符合條件的相鄰字組成 Bigram，並且運用模型以及意圖種子字來進行兩階段的意圖辭典擴充。最後採用了四種機器學習方法以及五個連續型的模型指標進行結果評估。

除了原始的研究數據外，本實驗中還刪除過度離散的票房值，並建立一個過濾後票房的實驗集進行比較。在對照後發現過濾後實驗集的模型評估結果值，普遍上明顯優於原始數據集，因此可以知道刪除差異太大的資料對於結果會產生一定的影響。比較原始數據的八個種子字以及敏感度分析的二十八個種子字，發現這兩種的評估結果並沒有太大的差異，大致上原始數據八個種子字的數值略優於二十八個種子字的數值，但是在決策樹(Decision Tree)當中，則有相反的結果。對比原始數據的全球票房以及敏感度分析的北美票房和北美首周周末票房，可以發現原始數據在 CC、RAE、RRSE 的值都明顯優於其他兩個敏感度分析的結果，就這個結論來看可以推敲可能是因為電影評論概括許多各國使用者的評論，若用這些評論來分析與北美票房的關係，可能會有些落差。

綜合以上不同的模型以及學習方法，可以發現到在所有的機器學習方法中，支援向量回歸(SVR)總是能在各個模型中都有最佳的模型評估結果，另外以 Unigram 為主的模型結果，像是: Google、Wiki 的預訓練模型以及用本實驗文本做的 Unigram 模型，普遍比 Bigram 的模型結果更好。從原始數據、過濾數據、以及三個敏感度分析的模型評估果彙整表當中可以明顯察覺，針對同一個模型，意圖辭典擴充兩次的數量會多於只擴充一次的，這可能是因為經由兩次的擴充過程，可以更廣泛的抓取到相關的意圖關鍵字。另外本研究也存在著一些關於電影數據蒐集的限制。近年

來，越來越多的熱門影片是由影音串流平台製作的，這些作品通常只在特定平台上發行，缺乏傳統的電影票房數據，而這對電影票房預測也帶來了一定的困難。

5.2 未來研究方向

以下針對本研究實驗提出未來改善的方向:

- **詞嵌入模型**

  本實驗使用的是 Word2vec 詞嵌入模型，但當前有不少文獻是參考更新的自然語言模型，例如: BERT (Bidirectional Encoder Representations from Transformers)。

- **Bigram 組成方法不夠彈性**

  本研究原先判斷 Bigram 的條件必需為緊鄰的兩個詞。建議可以放寬組成 Bigram 的條件，例如: "I will watch it again soon."，這句在原本的方法中並不會列為意圖字，但如果放寬 Bigram 的距離，就可以抓到"watch it again"意思也等於"watch again"。

- **Unigram 的意圖種子字不夠多元**

  八個意圖種子字的部分只有 Recommend 為 Unigram，在最後的模型結果以及附錄中可以發現，有些 Unigram 的模型即便只能針對 Recommend 進行擴充，但是結果卻很好。也因此建議可以提升 Unigram 做為意圖種子字的比例。

- **數據量不足**

  本研究的電影筆數只有 91 部，若又刪除掉離散值只剩下 81 部，而 Word2vec 的訓練文本有 618,125 筆評論，建議針對這兩項數據集都可以在做增加，以提升結果的準確率。

- **參考更多電影相關變數**

  在 IMDb 中有些變數是本次實驗沒有考慮的，例如:導演、演員、製片商等等，這些類別變數在資料筆數較少的狀況下是比較難處理的，但觀眾的確容易因為作品為知名導演、演員或片商所參與，而增加消費的動機，因此考慮加入這些變數，相信對票房會產生一定程度的影響。

參考文獻

Ahmad, Ibrahim Said, Azuraliza Abu Bakar, and Mohd Ridzwan Yaakub. 2020. 'Movie revenue prediction based on purchase intention mining using YouTube trailer reviews', *Information Processing & Management*, 57: 102278.

Ahmad, Ibrahim Said, Azuraliza Abu Bakar, Mohd Ridzwan Yaakub, and Mohammad Darwich. 2020. 'Beyond sentiment classification: A novel approach for utilizing social media data for business intelligence', *International Journal of Advanced Computer Science and Applications*, 11.

Breiman, Leo. 2001. 'Random forests', *Machine Learning*, 45: 5-32.

Chen, Zhiyuan, Bing Liu, Meichun Hsu, Malu Castellanos, and Riddhiman Ghosh. 2013. "Identifying intention posts in discussion forums." In *Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: human language technologies*, 1041-50.

Dhaoui, Chedia, Cynthia M Webster, and Lay Peng Tan. 2017. 'Social media sentiment analysis: lexicon versus machine learning', *Journal of Consumer Marketing*, 34: 480-88.

Ding, Xiao, Bibo Cai, Ting Liu, and Qiankun Shi. 2018. "Domain Adaptation via Tree Kernel Based Maximum Mean Discrepancy for User Consumption Intention Identification." In *IJCAI*, 4026-32.

Drucker, Harris, Christopher J Burges, Linda Kaufman, Alex Smola, and Vladimir Vapnik. 1996. 'Support vector regression machines', *Advances in Neural Information Processing Systems*, 9.

Galton, Francis. 1886. 'Regression towards mediocrity in hereditary stature', *The Journal of the Anthropological Institute of Great Britain and Ireland*, 15: 246-63.

Ghiassi, Manoochehr, David Lio, and Brian Moon. 2015. 'Pre-production forecasting of movie revenues with a dynamic artificial neural network', *Expert Systems with Applications*, 42: 3176-93.

Habib, Anam, Furqan Khan Saddozai, Anum Sattar, Aurangzeb Khan, Ibrahim A Hameed, and Fazal Masud Kundi. 2018. "User intention mining in bussiness reviews: a review." In *2018 5th International Conference on Behavioral, Economic, and Socio-Cultural Computing (BESC)*, 243-49. IEEE.

Hamroun, Mohamed, and Mohamed Salah Gouider. 2020. 'A survey on intention analysis: successful approaches and open challenges', *Journal of Intelligent Information Systems*, 55: 423-43.

Hutto, Clayton, and Eric Gilbert. 2014. "Vader: A parsimonious rule-based model for sentiment analysis of social media text." In *Proceedings of the international AAAI conference on web and social media*, 216-25.

Kabir, Md Rayhan, Faisal Bin Ashraf, and Rasif Ajwad. 2019. "Analysis of different predicting model for online shoppers' purchase intention from empirical data." In *2019 22nd International Conference on Computer and Information Technology (ICCIT)*, 1-6. IEEE.

Khattak, Asad, Anam Habib, Muhammad Zubair Asghar, Fazli Subhan, Imran Razzak, and Ammara Habib. 2021. 'Applying deep neural networks for user intention identification', *Soft Computing*, 25: 2191-220.

Li, ChenXing, YaJun Du, and SiDa Wang. 2017. "Mining implicit intention using attention-based rnn encoder-decoder model." In *Intelligent Computing Methodologies: 13th International Conference, ICIC 2017, Liverpool, UK, August 7-10, 2017, Proceedings, Part III 13*, 413-24. Springer.

Liu, Ting, Xiao Ding, Yiheng Chen, Haochen Chen, and Maosheng Guo. 2016. 'Predicting movie box-office revenues by exploiting large-scale social media content', *Multimedia Tools and Applications*, 75: 1509-28.

Maslowska, Ewa, Edward C Malthouse, and Vijay Viswanathan. 2017. 'Do customer reviews drive purchase decisions? The moderating roles of review exposure and price', *Decision Support Systems*, 98: 1-9.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. 'Efficient estimation of word representations in vector space', *ArXiv Preprint ArXiv:1301.3781*.

Quinlan, John R. 1992. "Learning with continuous classes." In *5th Australian joint conference on artificial intelligence*, 343-48. World Scientific.

Rashid, Ayesha, Muhammad Shoaib Farooq, Adnan Abid, Tariq Umer, Ali Kashif Bashir, and Yousaf Bin Zikria. 2021. 'Social media intention mining for sustainable information systems: categories, taxonomy, datasets and challenges', *Complex & Intelligent Systems*: 1-27.

Song, Tingting, Jinghua Huang, Yong Tan, and Yifan Yu. 2019. 'Using user-and marketer-generated content for box office revenue prediction: Differences between microblogging and third-party platforms', *Information Systems Research*, 30: 191-203.

Symeonidis, Symeon, Georgios Peikos, and Avi Arampatzis. 2022. 'Unsupervised consumer intention and sentiment mining from microblogging data as a business intelligence tool', *Operational Research*, 22: 6007-36.

Zeng, Daniel, Hsinchun Chen, Robert Lusch, and Shu-Hsing Li. 2010. 'Social media analytics and intelligence', *IEEE Intelligent Systems*, 25: 13-16.

Zhang, Li, Jun Li, and Chao Wang. 2017. "Automatic synonym extraction using Word2Vec and spectral clustering." In *2017 36th Chinese Control Conference (CCC)*, 5629-32. IEEE.

Zhang, Ning, Rong Liu, Xiao-Yang Zhang, and Zhi-Liang Pang. 2021. 'The impact of consumer perceived value on repeat purchase intention based on online reviews: by the method of text mining', *Data Science and Management*, 3: 22-32.

吳坤宗. 2010. '影評對電影續集票房影響之研究-以美國電影為研究對象', 國立暨南國際大學.

徐邦. 2015. '影響好萊塢電影票房銷售的決定因素之分析', 淡江大學.

附錄 A

附錄 A-1 本研究提出的意圖模型前十個擴充字

	recommend	**must watch**	**should watch**	**would watch**
1	would recommend(0.8166)	must see(0.9022)	should see(0.8501)	will watch(0.8167)
2	reccomend(0.8099)	should watch(0.6153)	will like(0.7491)	could watch(0.7684)
3	recomend(0.8013)	watch for(0.6136)	can watch(0.7334)	might watch(0.7004)
4	recommend for(0.7229)	should see(0.5183)	might like(0.7325)	would see(0.6882)
5	will recommend(0.7155)	can watch(0.5041)	will enjoy(0.7249)	may watch(0.6500)
6	recommend that(0.6879)	masterpiece(0.4898)	should like(0.7239)	would rewatch(0.6403)
7	recommand(0.6879)	recommend(0.4642)	can enjoy(0.7168)	will rewatch(0.6353)
8	can recommend(0.6191)	would recommend(0.4598)	might enjoy(0.7126)	can watch(0.6084)
9	recomment(0.6033)	should enjoy(0.4597)	will love(0.7114)	could rewatch(0.5822)
10	reccommend(0.6030)	will enjoy(0.4596)	should enjoy(0.7093)	will revisit(0.5796)

註:括號內的數字為擴充字與意圖種子字之間的相關性

附錄 A-2 本研究提出的意圖模型前十個擴充字

	watch again	**next time**	**can't wait**	**look forward**
1	rewatch(0.6778)	someday(0.4541)	cannot wait(0.6273)	excited(0.7074)
2	will rewatch(0.5923)	please(0.4401)	would love(0.4858)	eager(0.6004)
3	rewatche(0.5816)	hopefully(0.4381)	wait(0.4500)	would love(0.5856)
4	watch(0.5599)	asap(0.4026)	can wait(0.4152)	would like(0.5139)
5	watch that(0.5558)	should try(0.4003)	see above(0.4133)	wait(0.4958)
6	would rewatch(0.5216)	should stop(0.3782)	wait for(0.4049)	curious(0.4841)
7	watch over(0.5084)	pls(0.3702)	look forward(0.4006)	delighted(0.4764)
8	revisit(0.4924)	then(0.3579)	will wait(0.3982)	would compare(0.4656)
9	will watch(0.4880)	ya(0.3534)	could wait(0.3942)	glad(0.4622)
10	will recommend(0.4752)	will stop(0.3501)	eager(0.3894)	hope(0.4597)

附錄 B

附錄 B-1 Google 預訓練模型前十個擴充字

	recommend	must watch	should watch	would watch
1	recommended(0.7538)	null	null	null
2	recommending(0.6914)	null	null	null
3	recommends(0.6723)	null	null	null
4	consider(0.6032)	null	null	null
5	advise(0.5992)	null	null	null
6	propose(0.5328)	null	null	null
7	heartily recommend(0.5249)	null	null	null
8	recommendation(0.5209)	null	null	null
9	Recommend(0.4946)	null	null	null
10	Recommending(0.4724)	null	null	null

附錄 B-2 Google 預訓練模型前十個擴充字

	watch again	next time	can't wait	look forward
1	null	null	null	null
2	null	null	null	null
3	null	null	null	null
4	null	null	null	null
5	null	null	null	null
6	null	null	null	null
7	null	null	null	null
8	null	null	null	null
9	null	null	null	null
10	null	null	null	null

附錄 C

附錄 C-1 Wikipedia 預訓練模型前十個擴充字

	recommend	must watch	should watch	would watch
1	recommends(0.6986)	null	null	null
2	recommending(0.6889)	null	null	null
3	recommended(0.6652)	null	null	null
4	consider(0.6310)	null	null	null
5	prescribe(0.6022)	null	null	null
6	expect(0.6010)	null	null	null
7	advise(0.5973)	null	null	null
8	ignore(0.5911)	null	null	null
9	assure(0.5884)	null	null	null
10	approve(0.5790)	null	null	null

附錄 C-2 Wikipedia 預訓練模型前十個擴充字

	watch again	next time	can't wait	look forward
1	null	null	null	null
2	null	null	null	null
3	null	null	null	null
4	null	null	null	null
5	null	null	null	null
6	null	null	null	null
7	null	null	null	null
8	null	null	null	null
9	null	null	null	null
10	null	null	null	null

附錄 D

附錄 D-1 Phrases 模型前十個擴充字

	recommend	**must watch**	**should watch**	**would watch**
1	highly recommend(0.8264)	must see(0.8936)	should see(0.8152)	will watch(0.8200)
2	would recommend(0.8258)	everyone should see(0.6470)	will like(0.7500)	could watch(0.7770)
3	reccomend(0.7755)	should watch(0.6031)	might like(0.7434)	may watch(0.7137)
4	recomend(0.7498)	must must watch(0.5954)	can watch(0.7396)	might watch(0.6943)
5	will recommend(0.7417)	watch for(0.5871)	might enjoy(0.7297)	would see(0.6605)
6	strongly recommend(0.7261)	everyone should see(0.5277)	will enjoy(0.7242)	will rewatch(0.6471)
7	recommend for(0.6729)	can watch(0.5235)	should like(0.7157)	would rewatch(0.6376)
8	wholeheartedly recommend(0.6433)	see for(0.5097)	will love(0.7099)	will revisit(0.6255)
9	can recommend(0.6283)	masterpiece(0.5079)	may like(0.7012)	can watch(0.6207)
10	highly suggest(0.6188)	should see(0.5057)	should enjoy(0.6994)	could rewatch(0.6094)

附錄 D-2 Phrases 模型前十個擴充字

	watch again	**next time**	**can't wait**	**look forward**
1	rewatch(0.7077)	next next time(0.6367)	can't can't wait(0.6607)	excited(0.7198)
2	will rewatch(0.6034)	someday(0.5158)	can t wait(0.6016)	would love(0.5950)
3	rewatche(0.5775)	luck next time(0.4638)	cannot cannot wait(0.5912)	eager(0.5865)
4	would rewatch(0.5494)	please(0.4555)	cannot wait(0.5570)	eagerly await(0.5767)
5	watch(0.5476)	should try(0.4437)	via amazom(0.5546)	wait(0.5008)
6	watch that(0.5347)	ah(0.4248)	prime vod(0.5276)	highly anticipate(0.4943)
7	watch over(0.5178)	should stop(0.4209)	looke forward(0.5094)	would like(0.4920)
8	revisit(0.5098)	asap(0.4187)	will will I(0.5089)	delighted(0.4875)
9	will revisit(0.5090)	pls(0.4105)	would gladly(0.5046)	can't wait(0.4853)
10	recommend(0.5020)	plz(0.4079)	havn't(0.5018)	curious(0.4699)

附錄 E

附錄 E-1 Unigram 模型前十個擴充字

	recommend	must watch	should watch	would watch
1	reccomend(0.7855)	null	null	null
2	Recomend(0.7445)	null	null	null
3	recommand(0.6397)	null	null	null
4	advise(0.6201)	null	null	null
5	suggest(0.6151)	null	null	null
6	enjoy(0.5821)	null	null	null
7	reccommend(0.5552)	null	null	null
8	reccomende(0.5234)	null	null	null
9	recomment(0.5204)	null	null	null
10	watch(0.5016)	null	null	null

附錄 F-2 Unigram 模型前十個擴充字

	watch again	next time	can't wait	look forward
1	null	null	null	null
2	null	null	null	null
3	null	null	null	null
4	null	null	null	null
5	null	null	null	null
6	null	null	null	null
7	null	null	null	null
8	null	null	null	null
9	null	null	null	null
10	null	null	null	null