基於多模型深度學習架構之網路社群發文情緒分析

Multi-Model Deep Learning Frameworks for Emotion Analysis of Social Media Posts

張昭憲*
淡江大學資訊管理學系
jschang@mail.tku.edu.tw

許博翔
淡江大學資訊管理學系
408631066@gms.tku.edu.tw

摘要

線上社群發展快速，已成為現代人生活的一部分。然而，社群成員組成複雜，因意見分歧產生紛擾便時有所見。更由於網路發文之情緒傳播快速，若管理者無法及時發覺，便可能影響社群正常氛圍，甚至導致社群成員流失。因此，為維護社群長遠發展，首要之務便是了解成員情緒變化，並及早擬定對策。以人工定期審閱社群發文固然可行，但顯然緩不濟急，因此運用機器學習自動偵測社群發文情緒，便成為可行的解決之道。有鑑於此，本論文以情緒沙漏模型為基礎，配合深度學習發展了一套多層次多模型情緒偵測方法，以準確預測社群發文之複雜情緒。有別於傳統作法，提出之架構分為二層: 第一層建構了各種不同特質之二分類模型，分別以連續過濾與平衡過濾方式加以組合，以提升文章正負極性辨識準確性。在第二層，則提出以加權平均與投票方式組合多個二分類模型，進一步分類第一層之輸出，以產生更高解析之情緒值。為驗證提出方法之有效性，我們使用實際發文資料進行實驗。結果顯示，研究提出之多層次多模型架構確實優於傳統單一模型或直覺式多分類方法。當使用連續過濾與加權平均之組合架構時，更可獲得最佳之預測結果，改善幅度超過20%。上述成果除驗證提出方法之有效性，可做為網站管理之重要參考依據。

關鍵詞: 情緒偵測、情緒模型、深度學習、線上社群。

Abstract

Online communities are rapidly developing and have become a part of modern life. However, due to the complex composition of community members, disagreements often lead to unrest. Furthermore, the rapid spread of sentiment in online posts can affect the community's overall atmosphere and even lead to its gradual decline if managers fail to detect it promptly. Therefore, understanding member sentiment and developing effective countermeasures are paramount to maintaining the long-term development of these communities. While manual review of community posts is feasible, it is clearly insufficient. Therefore, using machine learning to automatically detect sentiment in community posts has become a viable solution. To address this, this paper develops a multi-layered, multi-model sentiment detection method based on the sentiment hourglass model and employs deep learning to accurately predict the complex emotions in social media posts. Unlike traditional approaches, the proposed architecture consists of two layers: the first layer constructs binary classification models based on various traits, combining them using continuous filtering and balanced filtering to improve the accuracy of identifying positive and negative polarity in articles. In the second layer, we propose combining multiple binary classification models using weighted averaging and voting to further classify the outputs of the first layer, generating a more refined sentiment score. To validate the effectiveness of our proposed approach, we conducted experiments using real-world publication data. The results demonstrate that our proposed multi-layer, multi-model architecture outperforms single models or intuitive multi-classification approaches. The combination of continuous filtering and weighted averaging achieves the best accuracy, with an improvement exceeding 20%. These results not only validate our proposed approach but also serve as an important reference for website management.

Keywords: Emotion Detection, Emotion Models, Deep Learning, Online Communities

壹、前言

現代人工作繁忙生活步調快速，使得傳統面對面溝通相當不易。然而，藉由線上社群(Online Communities)提供的溝通平台，讓人們無需出門也能聯絡感情、交換資訊，進而產生一種結合實際與虛擬的社交場域。在社群中，成員們可用虛擬身分與他人互動，大幅提升溝通方式的多樣性。時至今日，網路社群(如Facebook、Instagram、X、Line、TikTok等)早已成為人們不可或缺的生活重心。以年輕人慣用的TikTok平台為例，其2025年活躍用戶數已超過15億[19]。若不考慮重複註冊，已接近全球人口之20%。各類型社群提供的內容也不盡相同(如文字、語音、影片等)，伴隨分享訊息與回應，讓使用者在其中獲得不同程度的情緒抒發。

面對各類型線上社群，使用者有更多樣性的選擇。然而，也使社群管理面對更嚴苛的挑戰。首先，許多社群為增加用戶數，通常只需經過簡單認證便可加入。上述做法導致社群成員人數眾多且身分複雜，參與社群的目的更是難以管控。在網路的屏蔽下，成員們隨時可能帶給管理者難以處理的難題。例如，在選舉期間，公關公司成員可藉由在社群發文帶風向，影響特定候選人的支持度。又如，會員可能因不當敘述特定事件而導致網路霸凌，使當事人受到極大傷害[16][17]。管理者對於正面事件當然需予以鼓勵，但對於負面事件則需及早因應，以免社群成員受到不必要的情緒影響。由於網路社群競爭激烈，若不能瞭解用戶需求的變化，將使社群成長停滯，甚至日漸萎縮。因此，管理者在維持社群秩序之餘，需有更積極的作為，以維持社群的長遠發展。

由上述討論可知，現代社群管理的重要任務為能分析社群成員的活動，了解會員們的實際需求，以制定更合理的管理策略。為達成上述目標，可行的做法之一為預測成員的在社群中發文、回文的情緒，儘早發現社群氛圍的改變。過程中，不但需分析發文者的情緒，也需了解回文者的感受，才能制訂更全面的因應措施。情緒(emotions)是一種有意識的心理反應(conscious mental reaction)，表達人們的強烈感受，如喜樂(joy)、驚訝(surprise)、悲傷(sadness)等。會員們雖非面對面溝通，但可透過發文展現對特定人事物的情緒。管理者可藉由定時審閱社群中的發文與回文，進而了解成員們的情緒，以及其後續轉變。然而，面對數量眾多的社群發文，人工審閱的方式顯然曠日廢時且不切實際。當然，管理者亦可僅就熱門文章進行審閱以節省成本。但如此一來，便無法事發於未然，只能亡羊補牢避免事態擴大。

考量上述實務上需求，學者們紛紛提出各種自動情緒偵測的方法，以提供即時的情緒分析結果。考量多種基礎情緒，Bandhakavi 等人[5]以條件機率為基礎建發展複合式情緒偵測方法，並採用Part-of-speech, N-grams等技術來歸納文章屬性。Socher 等人[25]則透過深度學習方法-Recursive Neural Tensor Network進行文句的多等級情緒偵測，與傳統的Naïve Bayes、SVM等機器學習方法相較，可獲得更準確的預測結果。Suasanto 等人[27]以 SenticNet[22][23]為基礎，配合多種不同類型的情緒模型進行大規模實驗，以了解情緒模型的選擇是否與預測結果相關。值得注意的是，研究中雖考慮多維度的複雜情緒模型，但最終仍以文章正負極性(polarity)準確率做為評量指標。針對文章之正負情緒偵測，Chen[8]則以BiLSTM進行塑模，期能透過前後文分析，獲得更準確預測結果。Purba 等人[23]以三種基本情緒為基礎(angry, happy, sad)，對文章情緒進行三分類預測，分別考量以

Word Embedding配合深度學習技術，以及以TF-IDF配合傳統之機器學習方法。結果發現機器學習方法效果較佳，顯示傳統方法仍有其應用價值。Atandoh等人[4]則提出一套新的深度學習架構，整合位置嵌入與預訓練技術，並透過多通道卷積神經網路(MCNN)與注意力式bi-LSTM進行正、負情緒分類。提出之方法在IMDB等四個資料集上可獲得約85%~90%左右的準確率。Liu等人[18]則蒐集微博熱門話題，配合使用 多種情緒標籤(對應至情緒貼圖)，結合 BERT編碼器與neural topic模型進行情緒預測，結果顯示其準確率雖然不高，但仍優於現有作法。

除了基礎技術的發展外，情緒偵測也被應用在各種領域中。例如，Hu 等人[14]以Valence- Arousal 情緒模型為基礎[24]，透過搜尋關鍵詞相似度分析，嘗試為身心症患者找出適合閱讀的文章，以改善其情緒問題。Vadla等人[29]分析購物網站的產品評論，使用surprise, confuse等八種情緒標籤，並運用語意嵌入將評論轉換為向量，再使用BERT分類器進行情緒分類，以便將評論至設計維度，以找出產品的設計缺陷。Chen 等人[9]考量聊天室中慣用的表情符號，除文字發言外，也為表情符號建立情緒標籤。之後，再配合分群方法將訊息轉換為序列，以進行快速的情緒辨識。為避免負面發文對企業產生影響，Huang&Fang[13]則對社群中與特定公司相關之發文進行情緒分析，透過重構社群發文內容配合BERT建構辨識模型，以儘早標出對於公司不利的文章。為提升金融決策的準確性與即時性，Du等人[11]以金融新聞、財報資訊與社群發文為資料來源，採用BERT, FinBERT, LSTM 等分類器進行正面、負面、中性等三種情緒分析，結果顯示以專為金融語料設計之FinBERT表現最佳。為增強人機情感互動，Inoue等人[15]結合 BERT編碼器與 FastSpeech2 語音架構，提出階層式情緒分佈模型。配合使用Sad、Angry、Happy等五種情緒，可於語音合成步驟中進行強度控制，提升語音之情緒表達。有鑑於情緒偵測的實用性，Acheampong等人[3]探討了情緒分析情感偵測的進展、挑戰和機會，介紹相關的定義、任務與方法，並分析應用在不同領域和語言時之特點和困難。

由上述討論可知，情緒偵測對於現代社群管理極為重要。管理者若能事先了解會員情緒轉變，便可提早擬定因應對策，對社群進行積極管理。對於能引發正面情緒的發文予以鼓勵與推廣，對於引起紛爭甚至互相攻擊的文章，則透過社團規範予已制止，以避免過度發酵。唯有在情緒風險尚未擴散前即時介入，管理者方能在瞬息萬變的社群環境中維持秩序與穩定。針對上述主題，學者們雖已提出各種有效的作法，但我們發現相關研究方法仍有未盡之處，有以下明顯缺點有待改進: (1)人類情緒極為複雜，前人研究經常只考慮正負情緒(或數種基本情緒)，難以真實反映發文者的情感。應使用更精細的多維度情緒模型，才能提升結果之實用性。(2) 當採用高維度的情緒模型時，預測任務的複雜性會大幅提升，轉變為多元分類（multiclass classification）問題。然而，過去研究多以各種深度學習方法建構單一模型，其準確率常隨分類數量增加而顯著下降。為了提升預測結果之實用性，顯然有必要進一步發展更有效的模型建構方法。因此，本研究將致力於發展更有效的文章情緒偵測方法，期能同時顧及人類情緒之複雜性與分析結果之準確性，為上述問題提供可行的解決之道。

有鑑於此，本論文將以多維度情緒沙漏模型[27]為基礎，配合深度學習發展有效之多層次多模型偵測架構，以提供更準確之情緒偵測結果。有別於傳統作法，我們提出了一

套全新的雙層多模型偵測架構，將多分類任務分解，由多個二分類模型來達成。本研究提出之模型架構分為二層: 第一層為分流層(dispatch layer)，用於判別文章情緒之正負極性(polarity)。透過建構了各種不同特質的二分類模型，分別以連續過濾與平衡過濾方式加以組合，以提升極性辨識之準確性。第二層為解析層(resolution layer)，透過加權平均與投票方式組合多個二分類模型，進一步分類第一層之輸出，以提供更高解析度情緒值。為驗證上述方法之有效性，本研究以實際發文資料進行實驗。結果顯示，當解析層使用平均法與投票法時，結果確實優於傳統直覺式作法，並可產生超過10%的準確性改善。在分流層採用連續過濾式複合式架構時，亦可進一步產生更準確之極性分類結果。當第一層使用過濾法、第二層使用平均法時，可獲得最佳之偵測準確性，改善幅度高達20%。上述結果驗證本研究提出方法之有效性，確能提供更準確之多維度多情緒值之文章分析，可做為網站管理者重要參考依據。

貳、文獻探討與相關技術

為使後續討論更為順暢，以下將介紹與本研究相關之文獻、背景知識與術語，內容涵蓋情緒模型、深度學習技術與多模型偵測架構。

一、 情緒模型(Emotions Models)

為提供更細緻的預測結果，本研究將進行多維度多屬性值情緒偵測，這需配合適當的情緒模型，使機器學習方法有所依循。以下將介紹傳統的列舉式情緒模型、二維情緒模型，以及本研究使用的多維度情緒沙漏模型。

(一) 列舉式情緒模型

早期的情緒模型通常以基本情緒做為分類基礎，例如，被廣為採用的Ekman情緒模型[12]便將情緒分類為anger, disgust, fear, joy, sadness, surprise等六種。此後，為了描述更複雜的情緒，學者們更紛紛提出包含更多不同情緒的情緒模型，例如，Plutchik等人[20]將情緒分為acceptance, anger, anticipation等8種；Tomkins[28]提出的離散情緒理論更將情緒分為desire, happiness, surprise等9種。這些以基本情緒為主的模型雖然簡單易懂，但也產生以下缺點: (1)模型間中有許多用詞不同但意義類似的情緒，但卻存在「程度」關係，將其交互使用並不合適。例如，sorrow與sadness類似，但屬於更強烈情緒。 (2)某些情緒具有複雜意義，難以做為判別情緒之基礎。例如，surprise情緒便可能以正面(如意外收到同事禮物)或負面(如突然得知家人重病的噩耗)方式來展現[27]。

(二) 多維情緒模型

有鑑於人類情緒的複雜性，難以透過基本情緒來表達，因此學者們開始提出多維度情緒模型，以不同程度的感受組合出各種複雜情緒。例如，著名的Arousal-Valence情緒模型(參考圖1)便由Valence(鏈結方式)與Arousal(活力程度)二個維度所構成[23]。根據這二個維度便可產生四個情緒象限，分別為命名為Happy, Anxious, Depressed 與Content。人們可根據二個維度的值來決定目前的情緒狀態，例如，若Valence為較高的正值，

Arousal為輕度的負值，則可對應至Relaxed情緒。由於明顯較傳統列舉式情緒模型更符合實際需求，後續學者便紛紛提出以不同情緒維度為基礎的情緒模型[7][27][31]。

為提供高解析度更高的情緒預測結果，本研究將採用具有多維度多屬性值的情緒沙漏模型 (Hourglass model)。情緒沙漏模型是由Standford的Senticent團隊所提出 [7]，之後依照實務需求進行改良[27]。沙漏模型由四個情緒維度所組成:分別是Introspection, Temper, Attitude與Sensitivity。每一維度均有其代表之情緒詞彙，各情緒詞彙則代表在該維度下情緒展現之極性(polarity)與程度值(value)。沙漏模型之概念如圖2所示，各情緒維度所涵蓋的情緒詞彙與對應之情緒值則綜整於表1。據此，一篇文章的情緒便可以一情緒向量[v_i, v_t, v_a, v_s]來表示，其中v_i, v_t, v_a, v_s分別為其在各維度情緒對應之情緒值。舉例而言，若文章的情緒分析結果為[-3, -2, -1, +2]，則表示此文章展現由[哀痛(grief), 憤怒(anger), 不喜歡(dislike), 渴望(eagerness)]所組成之複合情緒，此結果顯然更接近人類所欲表達之複雜情感。鑑於上述特點，本研究將採用沙漏模型做為社群文章情緒偵測的基礎，透過將各維度之代表性情緒量化，再以向量方式組合，做為建立預測模型之根據。

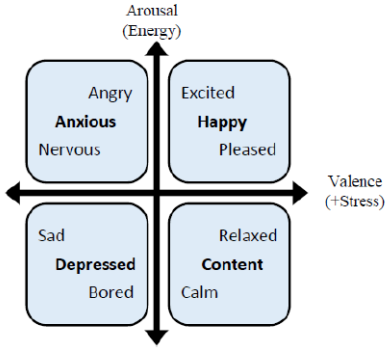


圖1: Arousal-Valence情緒模型[26]

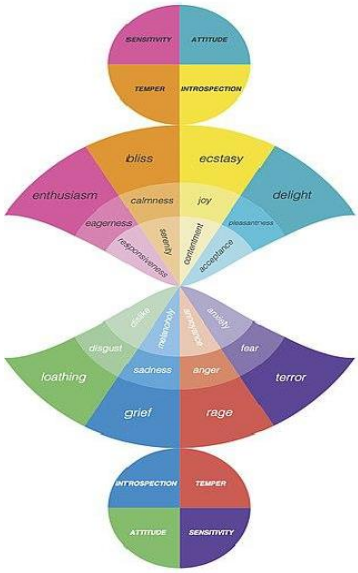


圖2: 情緒沙漏模型 [27]

表1: 情緒沙漏模型之維度情緒對應表 (整理自[27])

情緒維度	情緒類型(Emotions)					
Introspection	Ecstasy(狂喜)	Joy(喜悅)	Contentment(滿足)	Melancholy(憂鬱)	Sadness(哀傷)	grief(哀痛)
Temper	Bliss(幸福)	Calmness(寧靜)	Serenity(平靜)	Annoyance(煩惱)	Anger(憤怒)	Rage(暴怒)
Attitude	Delight(高興)	Pleasantness(愉快)	Acceptance(接受)	Dislike(不喜歡)	Disgust(厭惡)	Loathing(憎恨)
Sensitivity	Enthusiasm(熱情)	Eagerness(渴望)	Responsiveness(靈敏)	Anxiety(焦慮)	Fear(恐懼)	Terror(恐怖)
情緒值	+3	+2	+1	-1	-2	-3

二、 深度學習技術

本論文使用BERT(Bidirectional Encoder Representations from Transformers)[10]做為塑模方法，透過建構組合多個BERT模型，以產生多層次多階段之情緒偵測架構。BERT為Google於2018年所發布之先進自然語言處理技術，透過預訓練技術，讓模型可更快速應用於各相關領域。預訓練時，BERT採用大量未標籤化之文字資料做為輸入(包括Wikipedia(25億字)、BooksCorpus(8億字)等)，使其能理解各類型不同性質之文章。參考圖3之架構圖，BERT模型的主要結構由 Transformer編碼器 (encoder)堆疊而成，更透過雙向設計使編碼結果更容易處理文章前後文關係。換言之，訓練時會從目前token的前後上、下文學習，使能了解相同字詞在不同語境之不同意涵，以產生更接近人類之文章理解能力。由於預訓練模型已有一定之文章理解能力，使用者無需再次付出昂貴的模型訓練成本。透過新增的訓練資料和微調參數，便可解析不同類型的文章，並將編碼結果應用於不同學習任務中。在本研究中，BERT將被應用於多分類任務，藉以分辨評論文章在不同情緒維度之情緒值。透過BERT預訓練模型，便無需提供過於龐大之資料集，能有效節省總體開發時間。

如上所述，BERT是一種以變形器(Transformer)堆疊而成之深度學習架構。變形器是一種具有自注意力(self attention)機制的seq2seq學習模型[30]，架構是由一個編碼器(encoder)和一個解碼器(decoder)所組成(參考圖4)。當某一串列輸入至解碼器(decoder)後會被轉為詞向量(context vector)，並將目前的輸入字詞連同上下文傳入解碼器，以得到另一串列。自注意力機制的特點為記錄序列位置之間的關係，並根據輸入序列上下文來計算各個位置的權重，藉以加強或忽略某些字詞。因此，當以不同類型文章進行訓練後，相同的句子可能產生不同含意，使其產生與人類近似之語境感知能力。值得注意的是，序列的輸入與輸出長度可以不一致，因此亦可用於生成式AI等相關應用。

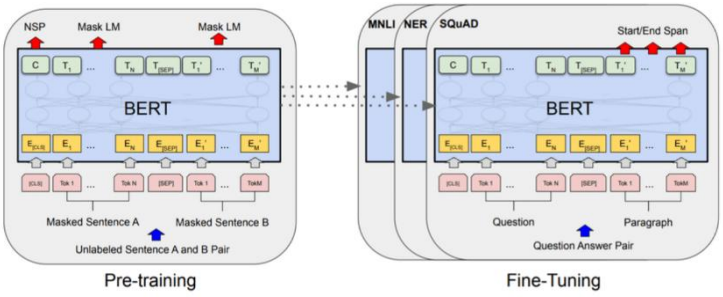


圖3: BERT 之 pre-training 與Fine-Tuning概念圖[10]

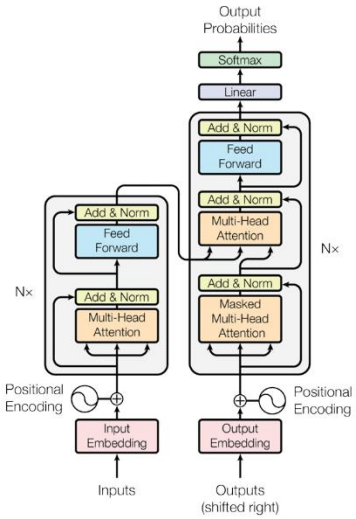


圖4: Transformer之模型架構[30]

三、 多模型多階段之分類架構

本研究將對文章進行多維度多情緒值之預測，為提升預測效能，因此設計了各種多層次多模型架構。在前人研究中，對於二分類任務也曾提出各種多模型預測架構。例如，針對線上拍賣詐騙偵測，文獻[6]以多模型為基礎設計了一套連續過濾流程。其主要概念為：由於許多待測帳號仍處於潛伏期，無法獲得足夠的詐騙特徵將其標出，因此設計了一套過濾式的偵測流程。參考圖5之流程架構圖，其中包含了M(100%), M(95%)等5個偵測模型，其中M(r%)模型乃使用者生命週期r%之交易資料所訓練之分類樹模型。透過連續運用五個不同時期交易歷史所建構之模型，便可以更準確方式，找出仍處於潛伏期的詐騙者。

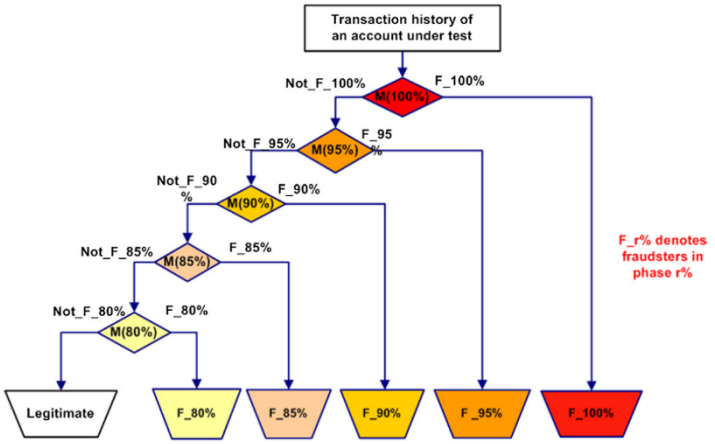
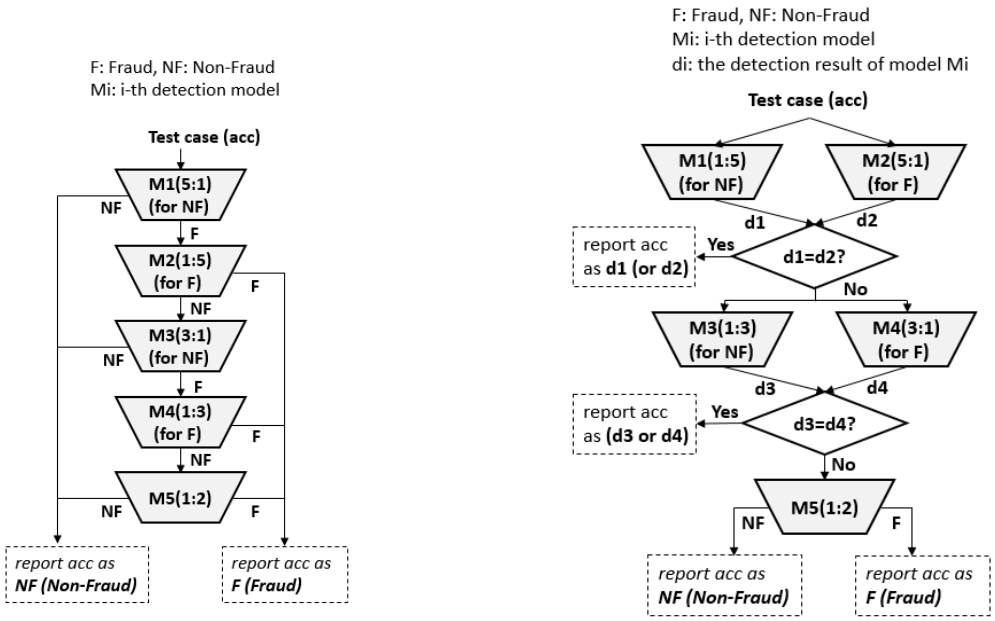


圖5：以連續過濾法進行線上拍賣之詐騙偵測[6]

同樣針對線上拍賣詐騙問題，文獻[1]則以資料配比調整為基礎，建構具備不同特徵的偵測模型，並透過多元方式進行組合，以提升偵測效能。參考圖6(a)，在此架構中，共包含5個二元偵測模型M(1:5), M(5:1), M(1:3), M(3:1), M(2:1)。他們發現，當以不同資料配比(Fraud:Non-Fruad)來進行訓練時，可產生不同效能之偵測模型，若能以適當方式加以組合，應可產生更佳之分類結果。由於各模型具有不同的偵測精度，因此逐次應用M(1:5), M(5:1), M(1:3), M(3:1), M(2:1)進行連續過濾，過程中若有任一模型(以其擅長之高精度分類)做出判定，則立即中止分類流程。為避免連續過濾法之決策過於武斷(因可隨時終止流程)，文章中也提出另一種多模型架構-平衡過濾法。參考圖6(b)，過程中將具有互補能力的模型做為一組，若二者之判別結果一致則終止流程，否則再進入下一階段的偵測。由上述說明可知，縱使對簡單的二分類任務而言，單一模型的效能仍有其極限。因此，為有效進行本研究所需之多分類任務，我們亦將以多模型架構為基礎，以不同方式加以組合，產生更準確的預測結果。



(a) 連續過濾法　　(b) 平衡式過濾法

圖6: 使用多模型架構進行線上拍賣之詐騙偵測[1]

參、以多模型為基礎之多維度情緒預測方法

為進行更積極的社群管理，管理者需要對社群文章進行情緒分析，以了解成員的情緒變化與整體社群氛圍動向。為此，本節將介紹本研究提出之以多模型為基礎之多維度情緒預測方法，以提供更有效、準確之文章情緒分析結果。

一、 問題陳述(Problem Statements)

在介紹提出方法細節之前，首先說明進行多維度多情緒值之情緒預測的複雜性。一般文章情緒分析僅考慮辨別正負情緒(positive/negative)，屬於二元分類問題。但當以情緒沙漏為基礎時，情緒的展現將以4個維度來代表，其輸出的可能性高達6^4種(參考圖7)，複雜度遠高於傳統的正負極性分析。若想獲得合理的準確性，顯然需發展更有效的預測方法。

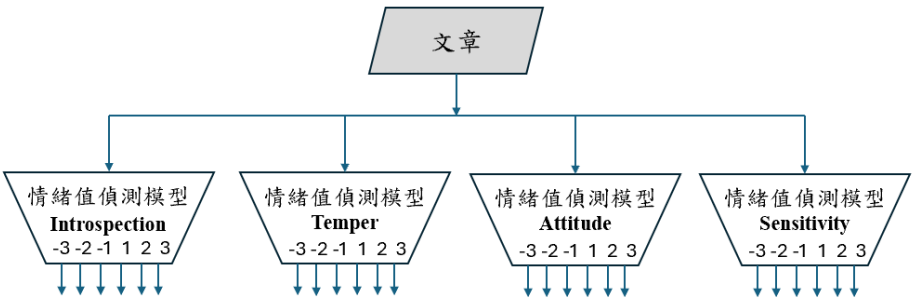


圖7: 將文章進行多維度多程度值之情緒預測

針對上述問題，最直覺方式為針對每一個情緒維度，均使單一分類模型(例如BERT分類器)來預測文章在該維度的情緒值。為使後續討論更為清楚，我將使用$M^{(n)}$來表示一個n分類的分類模型。參考圖8的架構，一個單一6分類模型$M^{(6)}$被用來做為多分類器，將待測文章分類為-3,-2,-1,1,2,3等六種情緒值。針對四個情緒維度，共只需四個模型來進行預測。此種作法相當直覺，因此我們將其稱為Naïve-C6 架構。此架構雖然簡單，但在執行多分類任務時，可能會遭遇以下困難:

(1) 首先，與二分類任務相較，多分類任務相對困難。若只使用單一模型來進行分類，容易因文章內容提供之「特徵」不足，難以區分太多類型的情緒值。根據表2(a)的前導實驗結果，縱使使用BERT來塑模，其準確率大約只有40%。

(2) 其次，由於模型的特性與其訓練集之各分類樣本數息息相關，若各分類之樣本數差異過大，也可能使模型效能受到影響。換言之，除非資料集特性能顧及所有分類之特性，否則分類數量越多，預測模型的效能可能逐次降低。相反地，若只有判別文章的正負極性(polarity)，則將可獲得極高的準確率。參考表2(b)的數據，當只考慮二分類時，其準確率可達92%。

很明顯地，直接使用單一多分類模型進行高維度情緒並不妥當。有鑑於此，本研究將發展各種多模型的預測架構，以有效提升預測準確率。

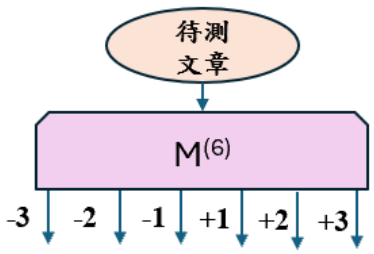


圖8: Naïve-C6 Framework-運用單一分類模型預測6種情緒維度值

表2: 使用單一BERT模型進行2分類與6分類任務之結果比較

(a)使用單一BERT模型進行6分類任務

6分類	Precision	Recall	F1-score	Support
-3	0.39	0.65	0.49	375
-2	0.38	0.25	0.30	375
-1	0.40	0.25	0.31	375
1	0.40	0.30	0.34	375
2	0.33	0.33	0.33	375
3	0.46	0.61	0.53	375
Accuracy			**0.40**	2250
Macro avg	0.39	0.40	0.38	2250
Weighted avg	0.39	0.40	0.38	2250

(b)使用單一BERT模型進行2分類任務

2分類	Precision	Recall	F1-score	Support
0	0.91	0.93	0.92	1125
1	0.93	0.91	0.92	1123
Accuracy			**0.92**	2250
Macro avg	0.92	0.92	0.92	2250
Weighted avg	0.92	0.92	0.92	2250

二、 雙層多模型之文章情緒預測架構

為改善使用單一模型的缺點，本研究設計了雙層多模型情緒偵測架構，將複雜分類任務分解，並以不同特質之模型來達成分解後的任務。參考圖9，本研究提出之基礎架構共分為二層:

(1) 第一層為分流層(dispatch Layer)，其目的為判別待測文章的正負情緒。由於此處屬於二元分類，理論上可獲得較高的準確率，因此將其置於第一層。之後，再跟根據文章正負情緒判別(Positive/Negative)，將文章分流至第二層-解析層。

(2) 第二層為解析層(resolution layer):其目的為判別文章之精確情緒值。若為正情緒文章則將進一步透過預測模型將其分類為+1,+2,+3等值；同樣地，負情緒文章則會被再細分為-1,-2,-3等值。

透過上述雙層架構設計，便可將原本的多分類任務拆解為二個階段，之後再在其中設計有效的多模型架構來配合。如此一來，便有機會改善傳統單一多分類模型準確率不佳的問題。值得注意的是，由於情緒沙漏模型有四個維度，圖9之雙層架構將會被應用在每個維度的情緒值預測上。

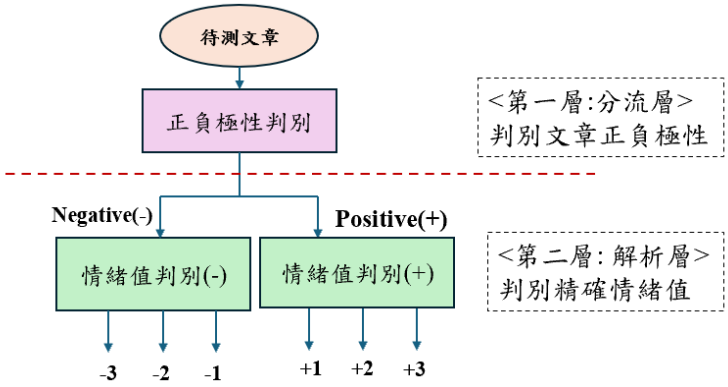


圖8: 本研究使用二階段雙層多維度多程度值情緒預測架構

根據上述架構說明，以下將介紹本研究提出的情緒預測方法。在所提出方法中，二層(分流層、解析層)均使用多模型架構來達成。為了使介紹過程更清楚， 以下將先介紹解析層的做法，內容涵蓋直覺法、投票法(Max-Count)及平均法(Averaged-Outcome)。之後，再介紹本研究建構之過濾式(Filtered-NP)與平衡式(Balanced-NP)多模型分流層架構。

三、 解析層之多模型架構設計(Multi-model frameworks for the resolution layer)

本節將介紹三種解析層架構- NP-C3、NP-MaxCount與NP-OAvg。其中， NP-C3為直覺式的簡易作法，NP-MaxCount與NP-OAvg則為複合式之多模型方法。 在本節中，第一層(分流層)均假設使用一個單一二分類預測模型$M_{NP}^{(2)}$，用以將待測文章分類為正、負情緒。此外，當將雙層架構命名為D-S時，表示其第一層(分流層)使用D架構，而第二層(解析層)使用S架構。例如，NP-C3代表第一層使用單一二分類模型$M_{NP}^{(2)}$，第二層則使用三分類模型。

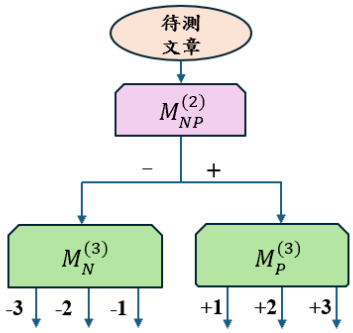


圖9: NP-C3多模型文章情緒預測架構(單維度)

(一) NP-C3 架構

以正負極性判別為前處理之架構，直覺的做法為將文章根據其極性再次進行分類。參考圖9，架構共分二層，第一層使用單一模型$M_{NP}^{(2)}$，對於文章的極性先進行分類，第二層則使用二個三分類模型$M_P^{(3)}$, $M_N^{(3)}$分別將正、極性文章再分為(+3, +2, +1)與(-3, -2, -1)。為簡明起見，以下將其稱之為NP-C3架構。上述架構之優缺點如下:

(a) 透過準確的二分類模型將文章分成正、負二種極性，之後再各分為 3 種情緒值，等同將六分類任務分為較低為維度的二分類與三分類模型組合。如此降維作法，將可能有助於提升總體準確。若將執行 c 分類的模型以 $M^{(c)}$來表示，則此架構即是令 $M^{(6)} \equiv 1*M^{(2)}+2*M^{(3)}$。

(b) 然而，使用$M_P^{(3)}$, $M_N^{(3)}$進行三分類任務時，仍會遭遇準確率下降問題。由於文章內容的用語特性，學習模型可能較無法分辨相近的情緒值(例如(-3, -2), (-1,-2), (1,2), (2,3)等)，導致總體準確率受到影響。參考表 3 之實驗結果，可發現$M_P^{(3)}$, $M_N^{(3)}$之總體準確率均不高(分別只為 0.52 與 0.45)。對於$M_P^{(3)}$而言，表 3(a)顯示情緒類別(+2)有極低的召回率(0.06)，顯然+2 類別大量被誤判為+1 或+3。 $M_N^{(3)}$雖無類似極端狀況，但各分類之精度與召回率均不理想。

考量上述問題，本論文將提出另二種不同的解析層多模型架構，進一步將三分類任務分解為由多個二分類模型來達成。藉由善用各模型的優點，將其輸出以合理方式整合，以產生更精確的預測結果。

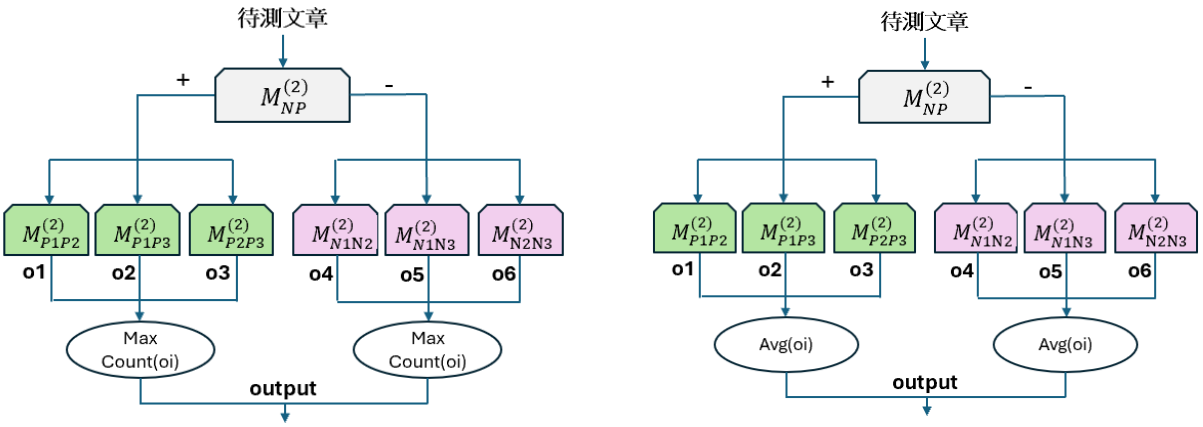
表3: 使用單一BERT模型進行3分類任務 (情緒維度: Introspection)

(a) 使用$M_P^{(3)}$進行正3分類 (+1,+2,+3) 之結果

3分類	Precision	Recall	F1-score	Support
+1	0.51	0.85	0.64	375
+2	0.38	0.06	0.10	375
+3	0.55	0.65	0.60	375
Accuracy			**0.52**	1125
Macro avg	0.48	0.52	0.45	1125
Weighted avg	0.48	0.52	0.45	1125

(b) 使用$M_N^{(3)}$進行負3分類 (-1,-2,-3) 之結果

3分類	Precision	Recall	F1-score	Support
-1	0.60	0.49	0.54	375
-2	0.35	0.53	0.42	375
-3	0.48	0.31	0.38	375
Accuracy			**0.45**	1125
Macro avg	0.48	0.45	0.45	1125
Weighted avg	0.48	0.45	0.45	1125



(a) NP-MaxCount 架構　　(b) NP-OAvg 架構

圖10: 本研究提出之二種解析層情緒預測架構

(二) NP-MaxCount 架構

以下介紹本研究提出之解析層架構- NP-MaxCount，其特點為透過多個二分類模型的投票結果，決定文章在特定維度的情緒值。參考圖10(a)，此架構的第一層使用$M_{P1p2}^{(2)}$將文章情緒分為正負二類，第二層則包含六個二分類模型，並分為二組: $M_{P1P2}^{(2)}$, $M_{P1P3}^{(2)}$, $M_{P2P3}^{(2)}$負責產生正情緒值， $M_{N1N2}^{(2)}$, $M_{N1N3}^{(2)}$,$M_{N2N3}^{(2)}$則用以產生負情緒值。其中，$M_{ab}^{(2)}$模型被用來將文章的情緒值歸類為a或b，如$M_{P1P2}^{(2)}$之輸出便只會是P1(+1)或P2(+2)，塑模時也僅用類別a與b的資料進行訓練。各模型之預測精度如表4所示，其中 $M_{P1P3}^{(2)}$與$M_{N1N3}^{(2)}$明顯優於其他四種(最高可達0.84)，此應導因於情緒值(+1,+3)、(-1,-3)較其他組合更容易分辨。但總體而言，表4各模型之精度明顯優於三分類模型(參考表3)。據此，當文章x被判別為正情緒時，透過$M_{P1P2}^{(2)}$, $M_{P1P3}^{(2)}$, $M_{P2P3}^{(2)}$等三組模型便可能產生以下八種輸出組合:

$$M^{P}{}_{out}(x) = \{ (o1,o2,o3) \mid o1 = M_{P1P2}^{(2)}(x),\ o2 = M_{P1P3}^{(2)}(x),\ o3 = M_{P2P3}^{(2)}(x) \} \quad (1)$$
$$= \{(1,1,2),(1,1,3),(1,3,2),(1,3,3),(2,1,2),(2,1,3),(2,3,2),(2,3,3)\}$$

檢視$M^{P}{}_{out}$可發現，1,2與3在可能組合中均出現8次。為了有效組合這三個模型的輸出，MaxCount採用文章在三個模型輸出中出現最多次的值為最終情緒值，也就是以投票方式決定最後情緒值。為更清楚進行說明，當文章極性為正時，其輸出組合以$M^{P}{}_{out}$表示，反之則以$M^{N}{}_{out}$表示。$M^{N}{}_{out}$亦有如下之8種不同組合，其特性亦與$M^{P}{}_{out}$相同。

$$M^{N}{}_{out}(x) = \{ (o4,o5,o6) \mid o4 = M_{P1P2}^{(2)}(x),\ o4 = M_{P1P3}^{(2)}(x),\ o6 = M_{P2P3}^{(2)}(x) \} \quad (2)$$
$$= \{(-1,-1,-2),(-1,-1,-3),(-1,-3,-2),(-1,-3,-3),(-2,-1,-2),(-2,-1,-3),(-2,-3,-2),(-2,-3,-3)\}$$

對於M_{out}中各種不同的組合，可能的MaxCount綜合輸出如表5所示。由表中上半部可知，當應用MaxCount至$M^{P}{}_{out}(x)$時，最後輸出情緒值為1,2或3之對應組合各有2種。對於(1,3,2)(2,1,3)組合，則因無法投票決定，則可選用任一情緒值做為輸出。同樣地，對於$M^{N}{}_{out}(x)$，MaxCount法亦有相同的結果(參考表3下半部)。

有關NP-MaxCount架構運作流程，其虛擬碼如表6所示。對於一篇待測文章x與特定情緒維度dim，先使用$M_{NP}(x)$產生其極性值polarity(line 2)。接著，根據極性值分別以p_models

或n_models中的三個模型(line 5-6)進行判別程度值的判別。若polarity為1，則以p_models中三個模型對文章p進行分類，並存入pdict中(line 8-10)。若polarity為-1，則將n_models中三個模型的判別結果存入ndict中(line 13-14)。最後，再以各模型的投票結果來決定p的情緒值ev(dim, p)。上述做法的優點如下: 雖然二分類模型的效能可優於三分類模型，但之間仍有效能差異，因此本架構透過多個二分類模型投票結果來決定最後情緒值，期能提升判別的準確性。

表4: 各種二分類模型之預測精度(情緒維度: Introspection)

Models	$M^{(2)}_{P1P2}$	$M^{(2)}_{P1P3}$	$M^{(2)}_{P2P3}$	$M^{(2)}_{N1N2}$	$M^{(2)}_{N1N3}$	$M^{(2)}_{N2N3}$
Precision	0.67(P1)	**0.82(P1)**	0.59(P2)	0.56(N1)	**0.84(N1)**	0.63(N2)
	0.63(P2)	**0.75(P3)**	0.55(P3)	0.54(N2)	**0.83(N3)**	0.60(N3)

表 5: 以MaxCount法組合多模型輸出以決定文章x之情緒值

$M^{P}_{out}(x)$	(1,1,2)	(1,1,3)	(1,3,2)	(1,3,3)	(2,1,2)	(2,1,3)	(2,3,2)	(2,3,3)
MaxCount輸出	1	1	{1,2,3}	3	2	{1,2,3}	2	3
$M^{N}_{out}(x)$	(-1,-1,-2)	(-1,-1,-3)	(-1,-3,-2)	(-1,-3,-3)	(-2,-1,-2)	(-2,-1,-3)	(-2,-3,-2)	(-2,-3,-3)
MaxCount輸出	-1	-1	{-1,-2,-3}	-3	-2	{-1,-2,-3}	-2	-3

表6: NP-MaxCount架構之運作流程虛擬碼

```
1   function NP_MaxCount
2      INPUT x: a text string to be classified
3      INPUT dim: the considered emotion dimension
4      INPUT MP1P2, MP1P3, MP2P3, MN1N2, MN1N3, MN2N3: sub-models
5      OUTPUT ev: the emotion value of x in dimension dim
6
7      polarity = MNP(x) # would be 1/-1 for positive/negative
8      pdict = {3:0, 2:0, 1:0}
9      ndict = {-3:0, -2:0, -1:0}
10     p_models = {MP1P2, MP1P3, MP2P3}
11     n_models = {MN1N2, MN1N3, MN2N3}
12     ev = 0
13     if polarity == 1:
14        for model in p_models:
15           pdcit[model(x)]+=1
16        ev = max(pdict, pict.get)
17     else: # polarity = -1
18        for model in n_models:
19           ndcit[model(x)]+=1
20        ev = max(ndict, nict.get)
21     return ev
```

(三) NP-OAvg 架構

除MaxCount法外，本研究也另提出之另一種解析層偵測架構-OAvg。參考圖10(b)，NP-OAvg使用之二分類模型數量與類型與NP-MaxCount相同，但整合時改將各模型之輸出

進行平均。舉例而言，若文章之極性為正，且透過$M^{(2)}_{P1P2}, M^{(2)}_{P1P3}, M^{(2)}_{P2P3}$產生之情緒值分別為o1, o2, o3，則文章之情緒值將為$(o1 + o2 + o3)/3$取四捨五入後之結果。若文章極性為負，則改用$M^{(2)}_{N1N2}, M^{(2)}_{N1N3}, M^{(2)}_{N2N3}$之輸出o4, o5, o6進行平均。此外，考量各二元模型有不同效能，亦可對其輸出值進行加權平均。若將對應至三個模型權重設定為r1:r2:r3，則文章x的情緒值便可能為$(o1 \cdot r1 + o2 \cdot r2 + o3 \cdot r3)/(r1 + r2 + r3)$或者 $(o4 \cdot r1 + o5 \cdot r2 + o6 \cdot r3)/(r1 + r2 + r3)$。舉例而言，針對文章$x$之極性為正且(o1,o2,o3)=(2,1,3)，若設定r1:r2:r3=1:2:1，則其最終之情緒值將為$(2 \cdot 1 + 1 \cdot 2 + 3 \cdot 1)/(1 + 2 + 1) = 1.75$(四捨五入後取2)。為求簡潔，本文後續將以 NP-OAvg(r1:r2:r3) 表示採用 r1、r2、r3 為模型權重之 NP-OAvg 架構。至於 OAvg 多模型架構的詳細運作流程，請參考表 7 所列之虛擬碼。因該流程已與前述說明相對應，此處不再重複敘述。

表7: NP-MaxCount架構之運作流程虛擬碼

```
1   function NP_OAvg
2      INPUT x, dim: the text to be classified, emotion dimension
3      INPUT r1, r2, r3: three integers for assigning model's weights
4      INPUT MP1P2, MP1P3, MP2P3, MN1N2, MN1N3, MN2N3: sub-models
5      OUTPUT ev: the emotion value of p in dimension dim
6
7      polarity = MNP(x) # would be 1/-1 for positive/negative
8      ev = 0
9      if polarity == 1:
10        ev = (MP1P2(x)*r1+MP1P3(x)*r2+MP2P3(x)*r3)/(r1+r2+r3)
11     else: # polarity = -1
12        ev = (MN1N2(x)*r1+MN1N3(x)*r2+MN2N3(x)*r3)/(r1+r2+r3)
13     return round(ev,0) # Round to the nearest integer
```

表8: MaxCount, OAvg(1:1:1), OAvg(1:2:1)等三種解析層方法之比較

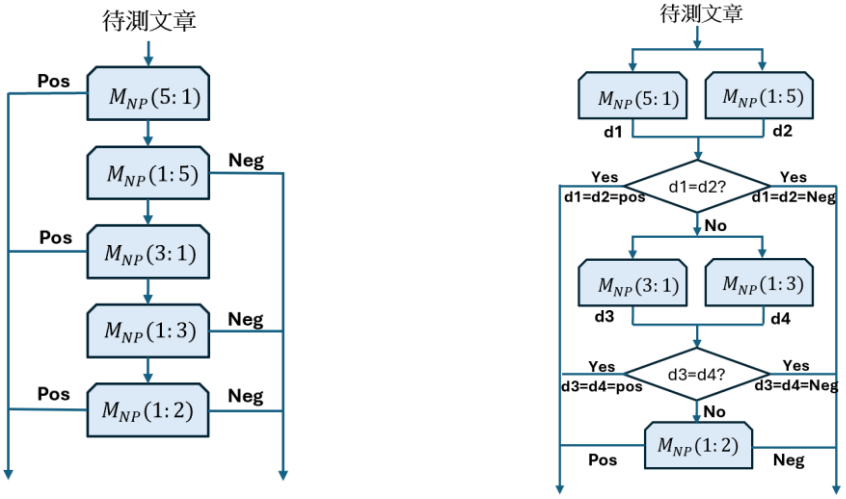
M_{out}	(1,1,2)	(1,1,3)	(1,3,2)	(1,3,3)	(2,1,2)	(2,1,3)	(2,3,2)	(2,3,3)
MaxCount	1	1	{1,2,3}	3	2	{1,2,3}	2	3
OAvg(1:1:1)	1	2	2	2	2	2	2	3
OAvg(1:2:1)	1	2	2	3	2	2	3	3

為了解OAvg法與上一節提出之MaxCont法的特質差異，表8列出MaxCount, OAvg(1:1:1)及OAvg(1:2:1)三種方法對於不同模型輸出組合(M_{out})所產生的情緒值。為簡明起見，此處只以正情緒值為例，其結果可直接推演至負情緒值。參考表OAvg(1:1:1)列可發現，與MaxCount相較，其輸出有2處差異。當在M_{out}為(1,1,3)時，MaxCout輸出為 1，而OAvg(1:1:1)則已平均方式輸出2；類似狀況也出現在(1,3,3)欄，此時MaxCout輸出3，而OAvg(1:1:1)則因平均而輸出2。對於(1,3,2),(2,1,3)二組組合，MaxCout因投票無法決定而任意選取其中一值，但OAvg法則確定選取中間值(2)。由上可知，由於取平均的緣故，OAvg(1:1:1)具有選取中間值(2)的傾向。此特性將使解析層在沒有明顯把握下，避免選取極端值1或3。當採用OAvg(1:2:1)時，則是將第二個模型$M^{(2)}_{P1,P3}$之輸出加權。如此一來，當M_{out}為(1,3,3), (2,3,2)時，其總體輸出便會成為3(與OAvg(1:1:1)的2不同)。換言之，因加

權緣故特別看重$M^{(2)}_{P1,P3}$的效能，當 $M^{(2)}_{P1,P3}$出現大的極端值時(3)，便會影響最後的綜合輸出。實際應用時，使用者可根據權重調整，改變三個子模型對於最後輸出之影響。

四、 分流層之多模型架構設計((Multi-model frameworks for the dispatch layer)

由於文章極性偵測為二分類任務，可獲得相當高之準確率，因此上一節提出之情緒預測架均以$M^{(2)}_{NP}$模型作為後續分類之前處理(第一層:分流層)。然而，我們發現若在前處理時便分類錯誤，則後續(第二層:解析層)的模型無論如何繁複，也無法產出正確的預測結果。例如，若$M^{(2)}_{NP}$模型之準確率為85%，則確定其最後的分類(6分類)結果不可能超過85%。在前人的研究中[1]發現，調整訓練時之各類型資料量的配比，可產生各種特性不同的模型。若能以合適方式加以組合，便可有效提升二分類任務的準確性。考量上述架構之有效性，本研究亦將運用類似概念，在分流層建構二種多模型架構 - Filtered-NP與Balanced-NP，以取代單一$M^{(2)}_{NP}$模型以提供更準確之文章極性判別。



(a)Filtered-NP連續過濾架構　　(b)Balanced-NP平衡過濾架構

圖11: 本研究建構之二種分流層架構(判別文章正負極性)

(一)Filtered-NP 連續過濾架構

參考圖11(a)，Filtered-NP架構共由5個二分類模型所組成，分別是$M_{NP}(5:1)$、$M_{NP}(1:5)$、$M_{NP}(3:1)$、$M_{NP}(1:3)$與$M_{NP}(1:2)$。其中，$M_{NP}(r1:r2)$表示此極性偵測模型之正、負訓練資料配比為r1:r2。受訓練資料配比的影響，各模型對不同分類任務因此展現不同效能(參考表9)。其偵測流程與各模型特質分述如下:

(a) 與正常之 $M_{NP}(1:1)$相比，$M_{NP}(5:1)$可獲得相當高之正面情緒預測精度，但對於負情緒之預測卻頗不準確。相反地，$M_{NP}(1:5)$模型可獲得極高的負情緒預測精度，但無法提供準確之正情緒預測。因此，本架構分別以 $M_{NP}(5:1)$與 $M_{NP}(1:5)$進行連續過濾，若文章經 $M_{NP}(5:1)$模型判定為正極性，則將其歸類為 Pos，不再進行後續流程。否則，則交由 $M_{NP}(1:5)$進行第二次偵測，若結果為負極性，則將其歸類為 Neg。當無法運用這二個模型來決定文章極性時，則交由 $M_{NP}(3:1)$與 $M_{NP}(1:3)$來判別。

(b) M_{NP}(1:3)和 M_{NP}(1:5)相較，其正面情緒預測精度略低，負面情緒預測卻較為準確。同樣地，M_{NP}(3:1)與 M_{NP}(5:1)亦有類似狀況。因此，在本階段將使用此二模型來協助分類上一階段無法判別極性之文章。

(c) 若經過上述四個模型仍無法決定其極性，則改用 M_{NP}(1:2)模型進行分類。由於 M_{NP}(1:2)對正、負情緒之判別效能較為均衡，因此被做為最後一個偵測模型。

當以過濾法進行正負極性判別後，便可將其串接至3.3節任一種解析層架構，以提升其總體準確率，藉此可產生Filtered-NP-C3、Filtered-NP-MaxCount、Filtered-NP-OAvg等架構。

表9: 以不同資料配比建構之極性判別模型效能比較

不同資料配比之模型	Precision (Neg)	Precision(Pos)	Accuracy	Data size
M_{NP}(5:1)	0.85	**0.95**	0.90	2250
M_{NP}(1:5)	**0.98**	0.80	0.87	2250
M_{NP}(3:1)	0.85	**0.94**	0.90	2250
M_{NP}(1:3)	**0.93**	0.90	0.91	2250
M_{NP}(1:2)	0.94	0.87	0.91	2250

(二)Balanced-NP 平衡偵測架構

Filtered-NP架構使用連續過濾方式進行文章情緒判別，其作法雖然有利於提升偵測效果，但各模型都只有一次參與判別的機會，可能使整體的判別過於武斷[1]。有鑒於此，本研究建構另一種文章極性判別方法-Balanced-NP架構，以更有彈性的方式來判別文章的正負情緒。參考圖11(b)之架構圖，Balanced-NP的第一層仍由5個模型所組成，分別為 M_{NP}(5:1)、M_{NP}(1:5)、M_{NP}(3:1)、M_{NP}(1:3)與M_{NP}(1:2)。不同於連續過濾法，此架構的判別只分為三個階段:

(1) 第一階段為使用二個對偶模型 M_{NP}(5:1)和 M_{NP}(1:5)分別預測文章之極性，若兩個模型的判別結果(d1=d2)相同，則直接進入下一層(解析層)之情緒維度值判別。若二者不同(d1≠d2)，則在將進入下一階段之極性判別。此階段判別的設計原理如下: 由於 M_{NP}(5:1)與 M_{NP}(1:5)分別對於正情緒與負情緒有極高之預測精度，若二者的判別結果不同，便無法決定採用那一個模型的輸出。因此，不同於連續過濾法，此時便會讓待測文章進入下一階段的判別流程。

(2) 在第二階段中，同樣使用二個對偶模型 M_{NP}(3:1)和 M_{NP}(1:3)分別預測待測文章之極性。如前所述，M_{NP}(1:3)模型對於負極性文章之判別精度雖不如 M_{NP}(1:5)，但對於正面文章之精度卻較高，M_{NP}(3:1)與 M_{NP}(5:1)也類似上述狀況。因此，在此階段若二個模型輸出相同(d3=d4)，則直接結束並進入下一層(解析層)。若仍無共識，則讓文章進入本層架構之最後一階段判別。

(3) 在第三階段，則以 M_{NP}(1:2)模型進行預測，並採用其輸出做為最後結果。

與Filterd-NP相同，Balanced-NP架構可與3.3節提出之各種解析層架構串接，藉此可產生Balanced-NP-C3、Balanced-NP-MaxCount、Balanced-NP-OAvg等多模型分類架構。由上述各節的介紹，可了解本研究提出複雜的雙層多模型架構之概念與特點，也明顯呈現其與傳統做法之差異。由於各方法各有特色，下一節將藉由實驗來驗證其效能，並了解各種設計方法之優劣。

肆、實驗結果與討論

為了驗證本研究所提出方法之有效性，本節採用實際發文資料集進行實驗。在比較過程中，將考慮不同的分流層與解析層架構組合，並探討不同設計在高維度情緒預測任務中的表現與效能差異。

一、 實驗設定

本研究使用前人研究中廣為採用的IMDB資料集[1]來進實驗，此資料集由史丹佛大學由權威影評網站IMDB擷取電影評論彙整而成。資料集共包含50,000筆影評資料，並被賦予正面(Positive)與負面(Negative)二種情緒標籤。其中，正面與負面資料筆數均為25,000筆。由於本論文提出之方法乃針對多維度多情緒值，因此需重新為每篇文章產生多維度之情緒向量標籤。為文章標記多維度情緒標籤之成本相當高，依照文獻 [2]中的實驗，多人耗費數月僅產生約3,600筆人工標籤。有鑑於此，本研究將採用生成式AI系統來標註多維度情緒標籤。為縮小生成式AI標籤與人工標籤之差距，我們分別使用Gemini Pro、GPT-3.5-turbo與GPT-4-turbo-preview等三種模型版本進行標註，最後再取用與人工標籤差異最小之AI模型。比較時，本研究採用文獻[2]所製作之Yelp多維度情緒資料集(共3665筆)做為基準，該資料集中之情緒標籤均為人工標註。Yelp Review為多國店家意見評論網站所提供之公開評論資料集，資料數量龐大，相當具有參考價值。

表10: 生成式AI標註情緒標籤與人工標籤之差異比較(評量指標: RMSE)

Dimension/AI Models	Gemini Pro	Gpt-3.5-turbo	Gpt-4-turbo-preview
Introspection	1.47	1.24	1.02
Temper	1.77	1.47	1.37
Attitude	1.72	1.24	1.11
Sensitivity	1.58	1.30	1.13
All(考量4個維度)	1.64	1.32	1.17

然而，在無任何提示狀況下，生成式AI系統並無法針對情緒沙漏模型產生情緒標籤。因此，標註前本研究會給予合適的提示前文，使其了解情緒模型的概念與意義。標註結果之比較如表10所示，其中數值為與人工標籤之RMSE(越小越好)。由表中可知，針對四種情緒維度，GPT-4-turbo-preview之表現最佳，平均RMSE僅1.17。因此，我們將採用GPT-4-turbo-preview為IMDB資料集進行情緒標註。本研究使用IMDB資料集中之25,000筆資料進行實驗(正負各12,500筆)，完成標註後便可將資料集之情緒標籤擴充至四個維度，每個維度具有六種不同情緒值(-3到3)。建立各種預測模型時，本研究採用Google提出之BERT技術，使用的版本為BERT-base-uncased。由於資料皆為IMDB之長篇評論，因此建立BERT模型時，均將文章Token長度上限設為500。本研究使用Python(3.10版)進行開發，開發環境為Google Colaboratory Pro。塑模時之訓練、驗證與測試資料比例分別為Train:Validation:Test = 0.64:0.16:0.2。

[1] https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews

比較各種方法之效能時，我們使用RMSE(Root-Mean-Square Error)而非準確性(Accuracy)作為評量指標。原因如下: 對於多分類任務而言，若實際值為-3，預測值為-2, -1, +1, +2, +3均屬預測失敗。但事實上，將預測值-2遠比+1準確。舉例而言，假設某維度之測試資料共有5筆，實際值為A=[-3, -2, 1, 1, 2]，而B=[-2, -1, 1, 3, 2]、C=[1, -1, 1, 1, -2]為二種不同模型M_B, M_C之預測值。以準確率而言，M_B, M_C的效能無差異(均是0.4=2/5)，但當考量RMSE時，則可得到不同的結果。由於RMSE(A,B)=$\sqrt{(1^2+1^2+0^2+2^2+0^2)/5}$ = 1.095, RMSE(A,C)=$\sqrt{(4^2+1^2+0^2+0^2+4^2)/5}$ =2.569，由此可了解在此次預測中，模型M_B的表現明顯優於M_C。

二、 多層次多模型情緒預測架構之效能驗證

本節將使用上述產生之IMDB多維度情緒資料集進行實驗，並比較第三節介紹之各類型情緒預測架構之效能。

(一) 解析層之各種架構效能比較

首先，我們將驗證多層次預測架構，是否有助於改善以單一模型進行複雜多分類任務之效能。此外，也將驗證各種解析層方法(C3, OAvg與MaxCount)能否提升預測準確性。為此，本節先以直覺式正負極性偵測方法做為分流層，再配合不同解析層方法進行實驗，結果如表11所示。由表中數據可知，傳統單一六分類模型(Naïve-C6)之預測RMSE為1.4558。而在多層預測架構中， NP-OAvg(1:2:1)架構可獲得最佳之RMSE (1.2809)，其次為別為NP-OAvg(1:1:1) (1.2889)與NP-Max-Count(1.3204)，最後則為NP-C3之1.3334。為進行更精細的比較，我們以單一六分類模型(Naïve-C6)做為基準點(baseline)，計算各種多架構之改善比例。參考表中之Imp%列，表現最佳之NP-OAvg(1:2:1)可獲得12%左右的改善，最差的NP-C3則有8.4%的改善。由上述結果可知:

(1) 本研究提出之解析層架構(OAvg與MaxCount)確實有助於提升準確性，平均可獲得11%~12%的改善，表現優於直覺式二階段方法 (NP-C3)之8.4%。當配合更精細之分流層架構時，當可進一步提升效能。

(2) 多層次多模型架構表現確實優於傳統單一多分類模型。縱使採用排名居末的NP-C3架構，亦可獲得8.4%的改善。

在表11中，除平均RMSE外(Avg列)，各單一情緒維度(Int, Tem, Att, Sen)之結果也有類似趨勢，說明提出方法具有良好的穩定性。此外， NP-OAvg(1:2:1)之優勢應導因於其提升準確性較高之$M^{(2)}_{P1P3}$, $M^{(2)}_{N1N3}$模型權重為50%，使其重要性高於$M^{(2)}_{P1P2}$, $M^{(2)}_{P2P3}$, $M^{(2)}_{N1N2}$, $M^{(2)}_{N2N3}$等其餘模型。

表11: 以單一正負極性模型($M^{(2)}_{NP}$)為前處理之各種解析層預測架構效能比較

RMSE	NP-C3	NP-OAvg (1:1:1)	NP-OAvg (1:2:1)	NP-MaxCount	Naïve-C6 (6分類)
Int	1.3233	1.2669	1.2636*	1.2671	1.3717
Tem	1.3284	1.2601	1.2540*	1.3339	1.5448
Att	1.2906	1.2849	1.2687*	1.3043	1.4605
Sen	1.3915	1.3439	1.3374*	1.3766	1.4462
Avg	1.3334	1.2889	**1.2809***	1.3204	1.4558
Imp%	8.4%	11.5%	12.0%**	9.3%	-
Rank	4	2	1	3	5

*在該情緒維度之最佳結果　**(1.4558-1.2809)/1.4558=12%

表12: 以Filtered-NP正負極性判別為基礎之預測架構效能比較

RMSE	Filtered-NP-C3	Filtered -NP-OAvg (1:1:1)	Filtered -NP-OAvg (1:2:1)	Filtered -NP-MaxCount	Naïve-C6 (6分類)
Int	1.3350	1.2439	1.2366*	1.2407	1.3717
Tem	1.2080	1.1207*	1.1322	1.1960	1.5448
Att	1.1270	1.0753*	1.0817	1.1193	1.4605
Sen	1.2450	1.1826	1.1718*	1.2482	1.4462
Avg	1.2228	1.1556	**1.1555***	1.2010	1.4558
Imp%	16.0%	20.62%	20.63%*	17.5%	-
Rank	4	2	1	3	5

*在該情緒維度之最佳結果

表13: 以Balanced -NP正負極性判別為基礎之預測架構效能比較(指標:RMSE)

RMSE	Balanced-NP-C3	Balanced-NP-OAvg (1:1:1)	Balanced-NP-OAvg (1:2:1)	Balanced-NP-MaxCount	Naïve-C6 (6分類)
Int	1.2889	1.2476*	1.2590	1.2664	1.3717
Tem	1.3540	1.2641*	1.2657	1.3431	1.5448
Att	1.2640	1.2685*	1.2757	1.2909	1.4605
Sen	1.3882	1.3386*	1.3403	1.3803	1.4462
Avg	1.3237	**1.2797***	1.2851	1.3201	1.4558
Imp%	9.07%	**12.10%***	11.73%	9.32%	-
Rank	4	1	2	3	5

*在該情緒維度之最佳結果

(二) 分流層之各種預測架構效能比較

本節將前一節之分流層架構(使用單一$M^{(2)}_{NP}$模型)，改為本研究建構之Filtered-NP與Balanced-NP架構，再進行情緒預測實驗。參考表12，其中數據為以連續過濾法Filtered-NP為正負極性判別法之各種預測架構效能比較。由表中可知，仍以Filtered-NP-OAvg(1:2:1)之預測結果為最佳(1.1555)，其餘方法之效能排名與表3類似，依序為Filtered-NP-C3，Filtered-NP-OAvg(1:1:1)與Filtered-NP-MaxCount。若以改善百分比(Imp%)為評量指標，可更清楚呈現Filtered-NP-OAvg組合可獲得超過20%的準確性提升。總體而言，Filtered-NP前處理架構，可獲約16~20%之效能改善。上述結果清楚驗證Filtered-NP-OAvg之多分類架構之有效性，確能有效提升傳統單一模型之效能瓶頸。表13則為使用Balanced-NP配合各種分流層方法之結果，其結果以Balanced-NP-OAvg(1:1:1)為最佳

(1.2797)，Balanced -NP-OAvg(1:2:1)次之(1.2851)，其餘排名與表4相同。然而，與Naïve-NP相較，Balanced-NP在改善百分比 (Imp%)方面並不太明顯(大約為9~12%)，原因應是其效能原本就不如連續過濾法所致[1]。

為比較各種分流層方法之效能差異，我們將上述各表之實驗結果彙整於表12。由表14(a)可知，當使用Filtered-NP為分流層時(第二列)，可分獲最最優之前四名(參考數字上的排名標籤)。當考量最佳之解析層方法OAvg(1:2:1)時，Filtered-NP獲得之RMSE(1.1555)較之NP-OAvg(1:2:1)(1.2809)和Balanced-NP-OAvg(1:2:1)(1.2851)，改善相當明顯。此外，當與C3、OAvg(1:1:1)與MaxCount等解析架構配合時，Filtered-NP亦可獲得最佳結果。上述結果顯示 Filtered-NP方法確能有效提升正負分類之準確性，也再次驗證以NP-OAvg(1:2:1)做為分流層之有效性。當以改善百分比(Imp%)呈現比較結果時，Filtered-NP-OAvg架構之優勢更為明顯。參考表14(b)，Filtered-NP-OAvg架構可獲得超過20%的改善，遠高於NP-OAvg(12%)與Balanced-NP-OAvg(11.73%)。總體而言(參考圖12之長條圖)，當與各種解析層架構配合時，Filtered-NP架構可獲約16%~20%之準確性改善，明顯優於Balanced-NP與Naïve-NP之9%~13%與 8%~12%。

綜合上述實驗結果可知，本研究提出之雙層多模型情緒預測方法能有效提升預測準確性，並獲得以下發現:

(a) 將單一 *n* 分類模型拆解為多個 2 分類或 3 分類模型，確實有助於提升預測準確性。
(b) 承(a)，以不同方式結合多個分類模型之輸出，能獲得更準確之情緒預測結果。
(c) 縱使在處理 2 分類任務時(如判別文章極性時)，改以多個特性不同的 2 分類模型來達成，仍可能產生更佳之預測結果。

表14: 以不同極性前處理方法獲得之情緒值預測結果比較

(a) 以RMSE做為評量指標

極性前處理/情緒值判別	C3	OAvg (1:1:1)	OAvg (1:2:1)	MaxCount	Naïve-C6 (6分類)
Naïve-NP*	1.3334	1.2889	1.2809	1.3204	1.4558
Filtered-NP	**1.2228 (4)**	**1.1556(2)**	**1.1555 (1)****	**1.2010(3)**	1.4558
Balanced-NP	1.3237	1.2797	1.2851	1.3201	1.4558

* 使用單一二分類模型$M^{(2)}_{NP}$做為分流層　** (n): n為該方法獲得之預測結果排名

(b) 以改善百分比(Imp%)作為評量指標

極性前處理/情緒值判別	C3	OAvg (1:1:1)	OAvg (1:2:1)	MaxCount	Naïve-C6 (6分類)
Naïve-NP	8.4%	11.5%	12.0%	9.3%	-
Filtered-NP	**16.0%(4)**	**20.62%(2)**	**20.63%(1)**	**17.5%(3)**	**-**
Balanced-NP	9.07%	12.10%	11.73%	9.32%	-

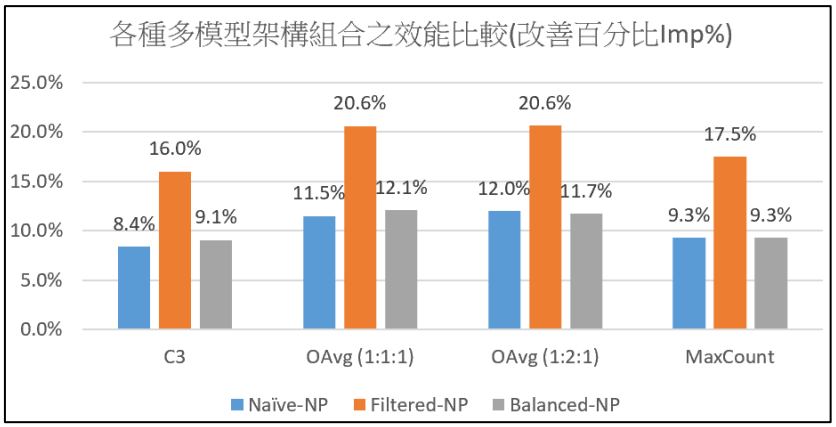


圖12: 各種雙層多模型架構之效能比較(以改善百分比Imp%為評量指標)

(三) 多模型架構之成本效益

表15為本研究使用的各種單一模型與偵測架構之塑模時間統計(單位:分鐘)，執行平台為Google Colab Pro，各模型均以BERT進行塑模，並選用T4 GPU(6MB L2 cache/16GB GDDR6)。由表中數據可知，傳統六分類模型 (Naïve C6)之塑模時間約為60 min。而NP-C3架構因需透過$M^{(2)}_{NP}$, $M^{(3)}_{N}$, $M^{(3)}_{P}$來達成，因此總時間約為114 min (=56+29+29)。當使用NP-MaxCount或NP-OAvg多模型架構時，由於需建構7個模型(參考圖10)，因此約需耗時174 min(=56+59+59)。而當採用最複雜的Filtered-NP或Balanced-NP架構時，在分流層需用5個2分類模型(參考$M^{(2)}_{NP}$的時間)，解析層需使用6個2分類模型，塑模時間預估最多為398 min (=56*5+59+59)。

與單一六分類模型相較，本研究提出之多模型架構雖然需較長的建構時間，但考量其帶來的準確率提升(可達20%左右)，在效能上仍具有顯著優勢。原因如下: 模型一旦建構完成，便可穩定重複使用於後續預測任務中，無需頻繁的重新訓練。換言之，初期建構成本雖高，但在長期應用中，便可展現出符合成本效益的實用價值。

表15: 本研究使用之各種模型架構之塑模時間統計 (情緒維度: Introspection)

(a)各種單一分類模型之塑模時間

Model	Naïve C6	$M^{(2)}_{NP}$	$M^{(3)}_{N}$	$M^{(3)}_{P}$	$M^{(2)}_{N1N2}$/ $M^{(2)}_{N1N3}$/ $M^{(2)}_{N2N3}$	$M^{(2)}_{P1P2}$/$M^{(2)}_{P1P3}$/$M^{(2)}_{P2P3}$
Time	60 min	56min	29min	29min	59min (20/19/20)	59 min(20/20/19)

(b)各種情緒偵測架構之總塑模時間

Model	Naïve C6	NP-OAvg/NP-MaxCount	Filtered-NP/Balanced-NP
Time	60 min	174 min	398 min
No. of models	1	7*	11**

*架構請參考圖10 **架構請參考圖11

伍、結論與未來工作

隨著網路的普及與相關技術蓬勃發展，網路社群早已成為了現代人生活不可或缺的一環。雖然忙碌，但人們藉由在社群分享訊息與回應，獲得不同程度的情緒抒發。然而，社群的氛圍瞬息萬變，稍有不慎便可能引發混亂，甚至因互相攻擊產生霸凌、騷擾等嚴重問題。凡此種種，如何控管社群總體情緒便成為管理者的重要課題。然而，社群成員的情緒變化快速，顯然無法以人工審閱方式來達成。因此，使用機器學習方法自動偵測社群發文的情緒，便成為可行的解決之道。

有鑑於此，本論文以多維度情緒沙漏模型為基礎，配合深度學習方法發展有效之多層次多模型偵測架構，以提升偵測之準確性。為此，我們提出了一套以文章極性偵測為前處理之雙層偵測架構: 第一層為分流層，用於判別文章情緒之正負極性；第二層為解析層，用於產生解析度更高之情緒值。在分流層中，我們建構了五個不同特質的二分類模型，分別以連續過濾(Filtered-NP)與平衡過濾(Balanced-NP)方式加以組合運用，以提升文章極性辨識準確性。在解析層中，我們提出MaxCount與OAvg二種架構，分別以加權平均與投票方式組合多個不同的二分類模型的輸出，以產生更準確的分類結果。為驗證提出方法之有效性，我們使用IMDB影評資料集進行實驗。結果顯示，當解析層使用OAvg與MaxCount架構時，結果確實優於傳統直覺式作法，並可產生超過10%的準確性改善。在分流層採用Filtered-NP複合式架構時，亦可進一步產生更準確之極性分類結果。尤其，當使用Filtered-NP-OAvg組合架構時，可獲得最佳之偵測準確性，改善幅度高達20%。上述結果驗證了本研究提出方法之有效性，確能提供更準確之多維度多情緒值之文章分析，可做為網站管理者重要參考依據。

有關文章之多維度情緒偵測，未來可針對以下問題進行更深入的探討：

(1) 為察覺情緒的細微變化，本研究將各維度情緒值分為六個等級，但在標記資料集情緒時常仍會有模棱兩可得情況。未來可導入更精細之情緒理論，重新評估各維度之情緒值，使分析結果更為實用。

(2) 由於實驗所用資料集之情緒標籤是由IMDB資料集擴充而成，因此缺少「無情緒」之情緒值。因此，未來亦可考慮將其加入資料集中，使預測結果更符合人類實際情緒反應。

(3) 本研究提出方法之預測效能提升，可能與將各模型輸出「模糊化」有關。然而，由於本研究採用啟發式策略(heuristic approaches)，透過觀察實驗結果來調整模型架構，若能在後續研究中發展更深入的相關性分析與檢驗方法，將有助於找出效能瓶頸，進一步提升總體準確性。

誌謝(Acknowledgement)

本研究感謝國科會專題研究計畫(NSTC 113-2410-H-032-053-)支持。This work is partially supported by National Science and Technology Council, Taiwan, R.O.C. under Grant no. NSTC 113-2410-H-032-053-.

參考文獻

[1] 張昭憲, 陳世軒。以模型融合為基礎之線上拍賣詐騙偵測。資訊管理學報，28(4)，2021 年: 頁 419-444。

[2] 田家豪(2021)，線上社群發文之複合情緒分析，淡江大學資管所，碩士論文。

[3] Acheampong, F. A., Chen W., Henry N.-M. (2020), "Text-based emotion detection: Advances, challenges, and opportunities", Engineering Reports, Volume 2, Issue 7, Jul 2020.

[4] Atandoh, P., Zhang, F., Al-antari, M. A., Addo, D., & Gu, Y. H., "Scalable deep learning framework for sentiment analysis prediction for online movie reviews", Heliyon, 10(10), 2024.

[5] Bandhakavi, A., et al., (2017) "Lexicon based feature extraction for emotion text classification," Pattern Recognition Letters, vol. 93, 2017: pp. 133-142.

[6] Chang, W.H. and Chang, J.-S. (2011)"A novel two-stage phased modeling framework for early fraud detection in online auctions", Expert Systems with Applications (38), 2011: pp. 11244-11260.

[7] Cambria, E., Livingstone, A., & Hussain, A. "The Hourglass of Emotions." *In Cognitive Behavioural Systems: COST 2102 International Training School, Dresden, Germany*, February 21–26, 2012: pp. 144-157.

[8] Chen, B., "Sentiment Analysis from Machine Learning to Deep Learning," *2021 International Conference on Electronic Information Engineering and Computer Science*, 2021: pp. 724-728.

[9] Chen, C.-H., Lee, W.-P., Huang, J.-Y., (2018) "tracking and recognizing emotions in short text messages from online chatting services," Information Processing and management, vol. 54, 2018: pp. 1325-1344.

[10] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K., "BERT: Pre- training of Deep Bidirectional Transformers for Language Understanding," last revised 24 May 2019 (v2), arXiv:1810.04805.

[11] Du, K., Xing, F., Mao, R., and Cambria, E., "Financial Sentiment Analysis: Techniques and Applications," ACM Computing Surveys, Volume 56, Issue 9, 2024: pp. 1 – 42.

[12] Ekman, P., & Friesen, W. V., "Measuring facial movement", Environmental Psychology & Nonverbal Behavior, 1(1), 1976: pp. 56-75.

[13] Huang, Z. and Fang, Z., "An Entity-Level Sentiment Analysis of Financial Text Based on Pre-Trained Language Model," *2020 IEEE 18th International Conference on Industrial Informatics (INDIN)*, 202: pp. 391-396.

[14] Hu, P., et al, "Valence-Arousal Analysis for Mental-Health Document Retrieval," 2017 *International Conference on Orange Technologies (ICOT)*, 2017: pp. 61-64.

[15] Inoue, S., Zhou, K., Wang, S., & Li, H., "Hierarchical emotion prediction and control in text-to-speech synthesis", *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024: pp. 10601–10604.

[16] Khatua, A., et al., "Tweeting in Support of LGBT?: A Deep Learning Approach," *Proceedings of the ACM India Joint International Conference on Data Science and Management of Data*, 2019: pp. 342–345.

[17] Kumar, A. J., Trueman, T. E., and Cambria, E., “A Convolutional Stacked Bidirectional LSTM with a Multiplicative Attention Mechanism for Aspect Category and Sentiment Detection,” Cognitive Computation volume 13, 2021: pp. 1423–1432.

[18] Liu, Y., Zhang, H., Wang, J., and Li, H., "LCSEP: A large-scale Chinese dataset for social emotion prediction to online trending topics", *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024: pp.1-5.

[19] Statista(2025, Feb), "Most Popular Social Worldwide by Number of monthly Active Users", https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users, Last Retrieved 2025/08/20.

[20] Plutchik, R., "The Nature of Emotions," American Scientist Vol. 89, No. 4, 2001: pp. 344- 350.

[21] Poria, S., et al., "Sentiment Data Flow Analysis by Means of Dynamic Linguistic Patterns," IEEE Computational Intelligence Magazine, Volume: 10, Issue: 4, 2015: pp. 26-36.

[22] Poria, S., Cambria, E., & Gelbukh, A., "Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis", *The 2015 conference on empirical methods in natural language processing*, 2015: pp. 2539-2544.

[23] Purba, S. A., Tasnim, S., Jabin, M., Hossen, T., & Hasan, M. K., "Document level emotion detection from bangla text using machine learning techniques", *In 2021 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD)*, 2021: pp. 406-411.

[24] Russell, J. A., (1980) "The circumplex model of affect," Personality and Social Psychology, 39(6), 1980: pp. 1161-1178.

[25] Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A. Y., & Potts, C., "Recursive deep models for semantic compositionality over a sentiment treebank", *In Proceedings of the 2013 conference on empirical methods in natural language processing*, 2013: pp. 1631-1642.

[26] Soleimaninejadian, P., Zhang, M., and Liu, Y. "Mood Detection and Prediction Based on User Daily Activities", *2018 First Asian Conference on Affective Computing and Intelligent Interaction* (ACII Asia), 2018.

[27] Susanto, Y., Livingstone, A. G., Ng, B. C., & Cambria, E., "The hourglass model revisited", IEEE Intelligent Systems, *35*(5), 2020: pp. 96-102.

[28] Tomkins, S., *Affect imagery consciousness: Volume I: The positive affects*, Springer publishing company, 1962.

[29] Vadla, M. K., Suresh, M. A. and Viswanathan, V. K., "Enhancing Product Design through AI-Driven Sentiment Analysis of Amazon Reviews Using BERT," Algorithms, 17(2), 59, 2024: pp. 1-17.

[30] Vaswani, A., "Attention is all you need", Advances in Neural Information Processing Systems, 2017: pp. 6000-6010.

[31] Whissell, C. M., *The dictionary of affect in language (Chapter 5)*, The measurement of emotions, Academic Press, 1989: pp. 113-132.