金融文本情感分析之探討

Transformer 模型與傳統方法之比較研究

李世勛

2025 年 6 月 6 日

摘要

在金融市場中，情感分析對於明智決策至關重要，它有助於預測趨勢、指導投資並評估經濟狀況。本研究比較先進的深度學習模型（如 Gemini、BERT 和 FinBERT）與傳統方法（如 SVM）在分析金融文本情感上的表現。我們首先採用零樣本分類，然後對各模型進行百次貝式優化微調。結果顯示微調後的模型，即使 FinBERT 在準確度上表現突出，但訓練調參時間長；而 Gemini 模型能大幅縮短訓練時間。這些發現突顯深度學習，尤其是 GPT 模型，在推進金融情感分類方面的巨大潛力，能為投資者和分析師提供寶貴洞見。

1 緒論

在日新月異的金融科技（FinTech）領域，從文字資料中洞察市場情緒——即所謂的「金融情感分析」（Financial Sentiment Analysis, FSA）——已成為一項不可或缺的關鍵技術。金融情感分析能夠深入揭示市場趨勢、輔助投資決策，並協助評估經濟指標的走向。透過精準判斷金融文本（例如新聞報導、財報、社交媒體討論）中所表達的情感，市場參與者可以更有效地制定策略，無論是預測股價波動，還是規劃投資組合，都具有實質的應用價值（Du 等人，2024）。考量到金融市場本身固有的高風險特性，運用先進的自然語言處理（NLP）技術，從海量的財經文件、新聞報導及社群平台的討論中，提煉出具有指標意義的情感資訊，其重要性不言而喻。

與一般領域的情感分析任務相比，金融情感分析面臨著巨大的挑戰。它需要解讀高度專業化的語言內容，其中包含專有術語、縮寫以及金融市場特有的詞彙。金融文本的語意往往較為模糊或細膩，其真實含義極度依賴上下文脈絡。舉例來說，一句在日常語境中可能被視為正面的話語，在金融情境下卻可能暗示著潛在風險或市場的不穩定。此外，金融市場的情緒時常受到總體經濟事件的影響，因此分析時必須考量時間的動態變化以及不同文本間的關聯性。這些特性使得金融情感分析成為一項極具挑戰性的工作，對分析模型的語意理解的細膩度與情境掌握能力都提出更高的要求。

本研究旨在比較經過微調（fine-tuning）的 Transformer 模型在金融情感分類任務上的表現，並與傳統的機器學習模型（如支持向量機 `SVM`、隨機森林、羅吉斯迴歸）進行直接的比較。具體而言，我們評估的對象包含 Gemini 模型、微調後的 `BERT` 模型、以及進一步針對金融領域微調的 `FinBERT` 模型。

透過比較這些不同模型的實際表現，本研究期望能對金融情感分析領域，以及基於 Transformer 的自然語言處理技術在金融應用上的研究做出實質貢獻。研究結果將提供關於微調大型語言模型（LLMs）在金融情感分類任務中效能的關鍵洞見，並突顯 Gemini 系列模型作為高效且準確分析工具的潛力。本專案所使用的程式碼與筆記本將會提供，以確保研究結果的可重現性。[1]|

2 文獻探討

傳統的金融情感分析方法主要依賴於情感詞典法和經典的機器學習模型，例如支持向量機（`SVM`）與羅吉斯迴歸（Du 等人，2023；Malo 等人，2014）。儘管這些方法在特定情境下展現一定的效果，但它們通常需要大量的特徵工程（feature engineering）以及針對特定領域的調整，才能處理金融語言中常見的技術術語、語境歧義，以及因經

[1]Colab 參考連結:
`https://colab.research.google.com/drive/1WBD18h4o_p2mnRzwatWQZmm05_9YEJH_?usp=sharing`

濟波動而產生的語意轉變。此外，基於規則的系統和傳統模型往往難以在不同的金融語境中進行有效的泛化，導致其在實際應用中的表現不盡理想。

深度學習技術的興起，特別是以 Transformer 為基礎的架構，徹底革新自然語言處理任務，金融情感分析亦在其列。這類模型透過在極大規模的文本資料上進行預訓練，能夠辨識複雜的語言模式，無需大量的人工特徵工程即可執行情感分類任務。其中一個代表性的成果是 Liu 等人 (2020) 以及 Y. Yang 等人 (2020b) 的研究，他們將 `BERT` 模型在 FiQA 和 PhraseBank 資料集上進行微調。這項工作催生 `FinBERT` 模型（GitHub，Y. Yang 等人，2020a）的誕生，有效地推動金融領域細緻情感分析技術的應用。這些 Transformer 模型能夠學習到文本的豐富特徵，使其能更好地捕捉對於準確金融情感分析至關重要的細微語意。

近期研究揭示基於 Transformer 的人工智慧模型在金融情感分析領域的潛力與挑戰。例如，Dmonte 及其團隊 (2024) 比較類似 GPT 的模型和基於 BERT 的模型。他們發現，儘管這些先進模型能有效解讀複雜的金融術語，但在資料稀少或情感表達模糊不清時，其表現仍會受到影響。另一方面，Fatouros 等人 (2023) 的研究則帶來令人鼓舞的結果。他們指出，若能巧妙地設計提示指令，ChatGPT 在外匯市場情感分析任務上的表現，相較於專為金融領域設計的 FinBERT 模型，能有高達 35% 的顯著提升。

3 研究材料與方法

為評估以 Transformer 為基礎的模型在金融情感分類任務中的效能，本研究對該特定領域的資料集上模型進行微調，以觀察模型在不同條件下的表現。參與比較的模型包括 Gemini、微調後的 `BERT` 及 `FinBERT`。傳統的機器學習模型，如支持向量機(`SVM`)、隨機森林及羅吉斯迴歸，則作為比較的基準。

儘管目前已有如 `LLaMA`、`Qwen` 等開源的大型語言模型，本研究選擇採用 API 形式的模型（如 Gemini）來執行任務。開源模型雖然具有較高的彈性與可控性，但其有效應用通常需要依賴多張高效能的 `GPU` 或 `TPU`，並且需要投入大量的時間與運算資源進行微調。相較之下，Gemini API 提供即時存取具有先進語言理解能力的模型之途徑，這些模型擁有優化的架構，並且在零樣本與少樣本任務中展現出穩定的效能。在資源與時間受限的情況下，此選擇提供一個更具實用性與可擴展性的解決方案。

本章節將詳細說明本研究所採用的資料前處理流程、模型訓練程序、評估指標以及實驗設計，藉此衡量各種方法在分類準確率與運算效率上的表現。

3.1 資料預處理

嚴謹的資料預處理是建構穩健機器學習模型的基石。此階段涉及清理原始資料、將其轉換為適合模型使用的格式，以及將資料切分為訓練集和測試集等步驟。

3.1.1 資料集描述

本研究選用「金融情感分析」（Financial Sentiment Analysis）資料集，此資料集旨在推動金融情感分析領域的研究進展（Sbhatti，2021）。該資料集整合兩個具代表性的資料來源：FiQA 以及 Financial PhraseBank。它以 `CSV` 檔案的形式提供，其中包含經人工標註情感的金融語句，並公眾授權方式在 Kaggle 和 Hugging Face 等平台上公開。

其中，Financial PhraseBank 部分最初由 Malo 等人於 2014 年建立，是其研究「Good Debt or Bad Debt: Detecting Semantic Orientations in Economic Texts」的一部分，該研究發表於《Journal of the Association for Information Science and Technology》（Malo 等人，2014；Malo & Sinha，2024）。他們研究中所使用的資料庫包含在 OMX 赫爾辛基證券交易所上市公司的英文新聞標題，這些標題是透過自動化網路爬蟲從 LexisNexis 資料庫下載而來。研究團隊從中隨機選取 10,000 篇文章，以涵蓋不同規模、不同行業的公司以及多元的新聞來源。依據 Maks 和 Vossen（2010）的方法，所有未包含可識別詞彙實體的句子都被排除。此步驟將樣本數量縮減至 53,400 句，每句至少包含一個或多個被識別出的詞彙實體。最後，從中隨機選取約 5,000 句作為整體新聞語料庫的代表，其情感標籤由 16 位標註者手動完成，平均的標註者間一致性達到 74.9%。

此資料集包含兩個主要欄位：「`Sentence`」欄位記錄金融文本內容，「`Sentiment`」欄位則標註情感類別（正面、負面或中性）。本研究選擇此資料集的原因在於其在 Kaggle 平台上的使用者評價極高（獲得 10/10 的評分），且內容完整性符合研究需求。

在進行資料前處理之前，我們先執行嚴謹的特徵工程，以深入瞭解資料的內在特性。原始資料集共包含 5,842 筆資料。「`Sentence`」欄位的最大字元數為 315，最小為 9，平均為 117。「`Sentiment`」情感標籤共分為三類：中性佔 3,130 筆、正面佔 1,852 筆、負面佔 860 筆。

儘管該資料集具有高度的研究價值，但情感類別的分布並不均衡，這可能導致模型在訓練時偏向於數量較多的類別，從而造成預測結果的偏差。為降低此風險，我們採用下採樣（down-sampling）策略來平衡各類別的資料分布，以確保模型能夠從更具代表性的數據中學習，提升學習效果。

3.1.2 資料集前處理流程

資料前處理階段採用結構化且可重複的流程，以確保資料品質並使其適用於預測模型。整體流程涵蓋多項步驟，包括缺失值處理、文本標準化、分層抽樣，以及將資

料集劃分為訓練集、驗證集與測試集。

首先，我們進行探索性資料分析，以檢視情感標籤的分布情形。此分析過程包含擷取情感類別及其對應的頻率，用以評估類別間的平衡程度。為此，我們呼叫專門的函式載入資料集，並統計各情感標籤的數量。接著，為維持資料的完整性，我們進行缺失值處理。透過預先定義的函式，篩選並刪除在「Sentence」與「Sentiment」等關鍵欄位中含有空值的資料列。該函式會將資料載入為 Pandas DataFrame 格式，檢測指定欄位的缺失情況，並移除含有缺失值的資料。

在完成資料清理後，我們進一步對文本資料進行正規化，以提升資料的一致性與模型的可解釋性。所採用的文本正規化方法包括將所有文本轉換為小寫、移除特殊字元、移除停用詞以及多餘的空白，確保整體格式的統一。與前述步驟相同，原始資料會另存新檔名，標準化後的版本則另存為 CSV 檔。

在描述性統計分析時，我們觀察到正面情感的樣本數量明顯多於其他類別，導致資料分布失衡。資料不平衡的問題會使模型在訓練時傾向於預測數量較多的類別，進而降低對數量較少類別的判別效能。為解決標籤不平衡的問題，本研究採用分層抽樣（stratified sampling）方法，此方法可以確保資料集中的各情感類別均保有相同的比例。最後，我們會將結果資料集打亂順序，以避免因順序造成的偏誤。

完成分層抽樣後，我們將資料集依照原始情感分布的比例劃分為訓練集、驗證集與測試集。首先，將 20% 的資料分配至測試集，其餘資料再進一步劃分為訓練集與驗證集。此過程中同樣透過分層抽樣，確保三個子集中的情感標籤皆呈現平衡分布。最終，這三組資料分別儲存為 train_set.csv、validation_set.csv 與 test_set.csv。最終的分布結果為：測試集 516 筆、驗證集 516 筆、訓練集 1548 筆，各情感類別在這三個資料子集中均勻分配。

同時需要考量的是，不同的機器學習演算法對於標籤格式的處理需求不盡相同。有些演算法可以直接處理字串型態的標籤，有些則須將標籤轉換為數值編碼。在本研究中，SVM 與 BERT 模型需要進行數值標籤轉換，而大型語言模型（LLMs）則可以直接處理類別型的字串標籤。對於維度較低的資料集，數值編碼是實用的選擇；而對於高維度資料，獨熱編碼（one-hot encoding）通常更為合適。本研究採用數值編碼方式，將情感標籤轉換為數值：負面轉為 0、正面轉為 1、中性轉為 2。我們在測試集資料中新增一個欄位來儲存這些數值標籤，以供後續模型評估之用。

3.2 模型微調與情感分析

完成資料準備後，即進入預測模型的訓練與微調階段。首先，我們評估傳統機器學習演算法的效能，包括支持向量機（SVM）、隨機森林與羅吉斯迴歸，並搭配廣泛的超

參數（hyperparameters）調整。同樣地，我們也對預訓練的 BERT 與 FinBERT 模型進行超參數微調，採用多種最佳化技術。此外，我們也針對 Gemini 模型進行微調。

起初，本研究曾考慮採用網格搜尋法（grid search）來尋找各模型的最佳超參數。然而，網格搜尋需要對所有可能的參數組合進行全面訓練，容易導致組合數量爆炸性增長，尤其是在同時處理多個 Transformer 模型與傳統演算法時更為明顯。為克服此限制，我們改採機率式的超參數調整方法，即貝氏最佳化（Bayesian optimization）。

本研究運用 Optuna 函式庫，針對每個模型執行 100 次試驗，廣泛搜尋超參數的組合。為提升訓練效率並避免過度擬合，我們整合提早停止（early stopping）與剪枝（pruning）技術。當模型在指定的訓練輪數（epochs）內效能未再提升時，即停止訓練。經過貝氏最佳化後，我們根據驗證集的準確率選出表現最佳的模型，並用於相同的測試集進行預測。在運算資源方面，本次研究分別採用 Google Colab Pro+ 中的 CPU 與 A100 GPU 資源，用於加速運算。其中，三種傳統機器學習演算法將使用前者（CPU），而 BERT、FinBERT 模型與 Gemini 模型將使用後者（GPU）。

3.2.1 SVM 情感分類與超參數優化

如前言所述，SVM 廣泛應用於分類任務，但需要進行大量的特徵工程、結合領域知識並細緻調整超參數，才能達到最佳效能。本研究採用 Optuna 框架進行貝氏最佳化，以微調 SVM 模型，使其適用於財經情感分類。整體流程在 Google Colab 平台上執行，並使用單一 CPU 完成。

本研究針對 TF-IDF 向量化（一種將文本轉換為數值特徵的方法）與 SVM 分類器進行全面性的超參數搜尋，以提升整體分類效能。詳細的超參數設定範圍如表 1 所示。其中停止詞設定部分，除了超參數設定中的 None 或"english"，也納入先前在正歸化使用到的停止詞作為超參數，查看是否為最佳參數。

超參數 (Hyperparameter)	搜尋範圍/值 (Search Range/Values)
TF-IDF Parameters	
max_features	5,000、10,000、15,000 或不設限
ngram_range	unigram (1,1)、bigram (1,2) 或 trigram (1,3)
stop_words	None (不移除) 或 "english" (移除英文停用詞)
SVM Parameters	
C	於 0.01 至 100.0 之間對數尺度取樣
kernel	'linear' (線性) 'rbf' (徑向基函數) 'poly' (多項式)
class_weight	可選擇設定為 “balanced” (以處理類別不平衡問題)
gamma	‘scale' 或 'auto'

表 1: TF-IDF 向量化與 SVM 分類器之超參數搜尋範圍

每次最佳化試驗都會建立一組結合 TF-IDF 與 SVM 的流程，在訓練集上訓練後，使用準確率與 hinge loss（一種損失函數）在驗證集上進行評估。所有試驗結果均被記錄下來，以便追蹤效能與辨識最佳的參數組合。整體方法的核心在於針對財經情感分類任務優化 SVM 分類器，並透過 Optuna 框架執行 100 次貝氏最佳化，以確保模型達到最佳效能。

表現最佳的模型出現在第 84 次試驗，其驗證集準確率達到 0.6899。該模型採用線性核函數，未限制 TF-IDF 的最大特徵數，使用 unigram 與 trigram，且未移除停用詞，正規化參數 C 約為 1.255。由於資料集中的各類別已經平衡，故未使用 class_weight 調整權重。

在測試集上，此 SVM 模型達到 0.6860 的準確率，且在各情感類別間表現一致。三個類別的 F1-score 皆介於 0.67 到 0.70 之間，顯示模型對於未曾見過的資料具有良好的泛化能力。

3.2.2 隨機森林模型優化與效能評估

延續相同的超參數調整方法，本研究在相同的訓練與驗證集上訓練隨機森林模型，以執行相同的分類任務。再次透過 Optuna 框架，廣泛搜尋超參數組合，以最大化驗證集的準確率。在所有試驗中，第 8 次試驗取得最佳結果，其驗證集準確率達到 0.6783。詳細的超參數設定範圍如表 2 所示。

超參數 (Hyperparameter)	搜尋範圍/值 (Search Range/Values)
Random Forest Parameters	
n_estimators	50 – 500
max_depth	10 – 100(step=10)
min_samples_split	2 – 20
min_samples_leaf	1 – 10
rf_max_features	'sqrt', 'log2', None
bootstrap	True, False
class_weight	None, 'balanced', 'balanced_subsample'

表 2: 隨機森林模型超參數搜尋空間

在超參數設定方面，最佳組合的 TF-IDF 向量器採用最大特徵數 5,000，n-gram 範圍為僅使用 unigram 與 trigram，且未進行停用詞過濾。在隨機森林分類器部分，最佳參數為：樹的數量 311 棵、最大樹深度 30、最小樣本分割數 11。此外，模型採用 log2 作為特徵選擇方式，未啟用 bootstrap 取樣，亦未進行類別權重調整。

在測試集上，模型達到 0.6531 的準確率，且在各情感類別間表現一致。三個類別的 F1-score 介於 0.60 至 0.68 之間，加權平均值為 0.64，顯示模型對於未曾見過的資

料具有良好的泛化能力。

模型中最重要的特徵多為常見詞彙與財經術語。前 20 個關鍵特徵包含 “the” “down”、“of the” 及 “up” 等詞，其中 “the” 的重要性最高，分數為 0.013，可發現大多特徵多為定冠詞與連接詞居多。另外，我們也設定移除停用詞後進行超參數設定並取得最佳組合進行預測，其驗證與測試準確率分別為 0.6473 與 0.5853，明顯較未設定移除停用詞時差。我們也可以觀察到，移除停用詞後的前 20 個關鍵特徵包含 "short"、"lower"、"decreased" 及 “long” 等詞。此結果顯示，文本中的常用詞彙，特別是與市場波動相關的字詞，對模型預測結果具有顯著影響。

3.2.3 羅吉斯迴歸模型優化與效能評估

羅吉斯迴歸模型亦採用與前述模型相同的超參數調整方法進行訓練，詳細的超參數設定範圍如表 3 所示。

超參數 (Hyperparameter)	搜尋範圍/值 (Search Range/Values)
Logistic Regression Parameters	
C	0.01 – 100.0(log scale)
class_weight	None, 'balanced'
max_iter	100 – 1000(step=100)
penalty_solver_combo	('l1', 'liblinear'), ('l1', 'saga'), ('l2', 'newton-cg'), ('l2', 'lbfgs'), ('l2', 'liblinear'), ('l2', 'sag'), ('l2', 'saga'), ('elasticnet', 'saga'), ('none', 'newton-cg'), ('none', 'lbfgs'), ('none', 'sag'), ('none', 'saga')
l1_ratio	0.0 – 1.0

表 3: 羅吉斯迴歸模型超參數搜尋空間

其最佳化結果如下：表現最佳的為第 75 次試驗，其模型所採用的最佳超參數組合包括：TF-IDF 向量器的最大特徵數為 10000，使用 1-gram 和 2-gram 特徵，未移除停用詞。正規化參數 C 最佳化為 2.781，並未啟用類別加權（balanced）。懲罰項與求解器設定為 'l1' 與 ‘saga'，最大迭代次數為 800。

在效能表現方面，最佳化共進行 100 次試驗，最終在驗證集上的評估結果顯示，模型準確率為 0.6822，loss 值為 0.7654。分類報告指出，模型在各類別間表現均衡，precision（精確率）、recall（召回率）與 F1-score 均接近 0.7，顯示整體效能穩定。在測試集上，模型達到 0.6802 的準確率，三個類別的 F1-score 介於 0.68 至 0.70 之間。分類報告顯示其表現在測試集上略低於驗證集。

3.2.4 BERT 模型微調與預測分析

隨著 Transformer 架構模型的興起，如 SVM 這類的傳統方法已逐漸被像 BERT 這樣的預訓練模型所取代。本研究亦採用與傳統模型相似的方法，對 BERT 模型進行微調。

本研究同樣採用貝氏最佳化，以高效搜尋 BERT 模型複雜的超參數空間。搜尋範圍如表 4 所示。

超參數 (Hyperparameter)	搜尋範圍/值 (Search Range/Values)
batch_size（批次大小）	4 – 64
learning_rate（學習率）	5×10^{-6} – 1×10^{-4}
epochs（訓練輪數）	2 – 10
optimizer_type（優化器類型）	Adam, AdamW
weight_decay（權重衰減）	1×10^{-6} – 1×10^{-2}
warmup_ratio（預熱比例）	0 – 0.2
dropout_rate（丟棄率）	0.1 – 0.5

表 4: BERT 模型超參數搜尋範圍

本實作採用 Hugging Face (2024) 提供的 BERT base uncased 模型，並針對序列分類任務進行設定。dropout_rate 則依最佳化試驗結果進行調整。每次試驗的訓練流程包括：執行斷詞處理並進行適當的補齊與截斷；透過自定義的資料集（Dataset）與資料載入器（DataLoader）類別準備批次資料；在訓練過程中追蹤損失並更新梯度；每個 epoch 結束後於驗證集上進行評估；並根據驗證準確率趨勢採用提早停止（early stopping）機制，以避免過度擬合。

表現最佳的為第 82 次試驗，其驗證集準確率為 0.760。對應的超參數組合為：batch_size 為 4、learning_rate 為 2.49×10^{-5}、訓練 7 個 epochs、採用 Adam 優化器、weight_decay 為 0.001、warmup_ratio 為 0.054，dropout_rate 為 0.263。在測試集上，模型達到 0.7326 的準確率，三個類別的 F1-score 介於 0.70 至 0.77 之間。分類報告顯示其表現在測試集上略低於驗證集。

3.2.5 FinBERT 模型微調與預測分析

本研究亦以相同流程應用於 FinBERT 預訓練模型。FinBERT 是專為財經領域優化的 BERT 變體，預先在財經文本上進行訓練，並針對情感分類任務進行調整。

超參數調整涵蓋多項設定，搜尋範圍如表 5 所示。

表現最佳的超參數組合於驗證集上達成 0.833 的準確率。該組合包含：batch_size 為 16、learning_rate 為 9.87×10^{-5}、訓練 2 個 epochs、採用 AdamW 優化器、

超參數/設定	範圍/值
可調整超參數 *(Tunable Hyperparameters)*	
`batch_size`（批次大小）	4, 8, 16, 32, 64
`learning_rate`（學習率）	$5 \times 10^{-6} - 1 \times 10^{-4}$
`epochs` (訓練輪數)	2 – 10
`optimizer_type` (優化器)	`Adam`, `AdamW`
`weight_decay`（權重衰減）	$1 \times 10^{-6} - 1 \times 10^{-2}$
`warmup_ratio`（預熱比例）	0.0 – 0.2
`dropout_rate`（丟棄率）	0.1 – 0.5
固定設定 *(Fixed Settings)*	
`max_sequence_length` (最大序列長度)	512
`CUDA` 加速	啟用 (若可用)

表 5: FinBERT 模型超參數調整設定詳情

`weight_decay` 為 8.48×10^{-6}、`warmup_ratio` 為 0.014，`dropout_rate` 為 0.109。在測試集上，模型達到 0.8333 的準確率，三個類別的 `F1-score` 介於 0.82 至 0.85 之間。分類報告顯示其表現與驗證集相當。

3.2.6 Gemini 模型微調與預測分析

與前三個傳統機器學習及 `BERT-based` 模型，需要透過 Google Colab 上的 `CPU/GPU` 進行訓練不同，Gemini 模型通常為封閉原始碼，故其微調與預測皆透過 Gemini API 執行。為此，本研究實作專用類別，用於管理 API 連線、建立提示語（prompt）、處理模型回應，並產出微調所需的 `JSONL` 格式檔案。

為使大型語言模型（LLM）生成具有意義的文字，需要設計結構合理的提示語，以有效地引導模型、優化 `token` 的使用並降低成本。為達此目的，本研究採用既有的提示語工程策略（K. Zhang 等人，2024），其核心方法主要涵蓋以下兩個方面。

首先，在提示語的內容設計上，會強調「與特定模型無關的通用性」。提示語採用通用格式撰寫，不依賴特定的模型架構，具備跨模型的適用性。此設計強調任務目標的明確傳達，並結合語境相關的指示，有助於與多種大型語言模型順利整合。其次，為了讓模型產出的內容更易利用，會著重於「結構化的輸出格式」。為同時滿足人類閱讀與機器處理的需求，回應格式遵循標準化的編碼與可存取性原則。輸出內容採用 `JSON` 標準結構，具備邏輯性的排列，便於自動化系統後續處理與整合。

本次測試採用的模型皆為 Gemini 1.5 Flash，此模型是 Gemini 系列中的輕量版模型，專為高效處理而設計，能夠在保留高效能的同時減少資源需求，在長內容檢索中具備高度的準確度。經過多次在各類大型語言模型中反覆測試與調整後，最終設計出一組可穩定產生目標格式輸出的提示語。此優化結構同時提升人類的理解度與模型的

運行效率。最終版本如清單 1 所示。

```
conversation.append({
    "role": "user",
    "content":
        "You are an AI system focused on evaluating sentiment in financial
   texts. Your role is to assess each sentence related to finance and
   determine whether its tone is positive, negative, or neutral. Return the
    result in the specified JSON format, and do not include any additional
   commentary."
        {"Sentiment": "sentiment_tag"}}. \nFinancial sentence: ..."
})
```

Listing 1: 用於零樣本推論的模型無關提示語結構

使用最終版本的提示語後，我們進行一系列迭代測試，分別引導 Gemini 模型對測試集進行情感預測。此階段的預測完全基於模型在預訓練期間所學到的知識，未進行任何針對此任務的專屬微調。模型執行的是零樣本（zero-shot）分類，依據句子的語意判斷其情感傾向並指派情感標籤。預測結果以 JSON 格式輸出，透過專用方法處理後儲存於測試集檔案的獨立欄位中；同時，亦記錄每筆預測的回應時間，並分別存放於各模型對應的欄位中。

為進行微調作業，本研究透過遍歷訓練集與驗證集產生兩個 JSONL 格式的檔案，內容為結構化的提示-回應（prompt-response）配對，並儲存於相同的檔案中，如清單 2 所示。檔案準備完成後，上傳至 Gemini 使用者介面，並選擇欲微調的目標模型以啟動訓練流程。

```
{"messages": [
  {"role": "system", "content": "You are an AI assistant specializing in
    financial sentiment classification."},
  {"role": "user", "content": "You are an AI assistant specializing in
    financial sentiment classification. Your task is to analyze each
    financial sentence and classify it as negative, positive, or neutral.
    Provide your final classification in the following JSON format without
    explanations: {\"Sentiment\": \"sentiment_tag\"}}. \nFinancial sentence:
     ..."},
  {"role": "assistant", "content": "{\"Sentiment\": \"neutral\"}"}
]}
```

Listing 2: 用於 JSONL 微調檔案的提示與完成配對

儘管 Gemini 提供讓使用者在微調過程中手動設定超參數的彈性，但本研究基於時間與資源限制，選擇採用平台預設的超參數組合。預設設定包含訓練 5 個 epoch（訓練

週期）、批次大小（batch size）為 16，以及學習率倍率為 1。此組合雖為通用設定，但極有可能是 Gemini 最佳化流程所產出的結果，能作為多數應用情境下的穩定起點。

由於免費版服務的限制，訓練參數一次僅能允許使用 100 筆資料進行模型建構。因此，本次研究在此架構下進行訓練。若對於生產等級的應用或對效能要求更高的研究，仍建議針對特定的資料集與目標，進行專屬的超參數最佳化，以獲得更佳的表現。微調程序完成後，使用新訓練完成的模型對測試集進行預測。與先前的預測流程相同，將情感分類結果與模型回應時間一併記錄於測試集檔案中，以供後續分析使用。

4 實驗結果

第三章詳細說明本研究在財經情感分析中所採用的 GPT（此處指 Gemini）與 BERT 模型的實作與微調方法，並介紹針對 SVM、隨機森林與羅吉斯迴歸模型所進行的超參數調整策略。本章節將針對這六種模型進行綜合比較分析，重點涵蓋其效能指標、預測時間與運算成本等面向。各模型在微調前後的預測評估指標統整如表 6 所示。

模型名稱	準確率	精確率	召回率	F1 分數
ft:gemini1.5flash	0.8198	0.8300	0.8200	0.8200
base:gemini1.5flash	0.8120	0.8133	0.8133	0.8100
ft:finbert	0.8333	0.8333	0.8333	0.8333
ft:bert	0.7326	0.7333	0.7333	0.7367
t:svm	0.6860	0.6933	0.6867	0.6867
t:random-forest	0.6473	0.6533	0.6467	0.6433
t:logistic-regression	0.6802	0.6900	0.6800	0.6833

表 6: 各模型效能指標比較

4.1 Gemini 1.5 Flash 零樣本評估

Gemini 1.5 Flash 首先進行零樣本（zero-shot）評估，即在未經任何針對財經領域的微調情況下，直接執行財經情感分類任務。其表現如下：準確率為 0.812，精確率、召回率與 F1 分數均約為 0.81。上述結果顯示，即使在零樣本條件下，預先在通用語言資料上訓練的大型語言模型，仍能有效地推論出與財經領域相關的語境和特徵，並用於情感分類任務。

4.2 模型微調後效能評估

在零樣本評估之後，我們進一步針對 Gemini 1.5 Flash 在財經情感資料集上進行微調。無論是微調前或微調後，其表現均顯著優於傳統模型。Gemini 1.5 Flash 的準確

率由零樣本的 0.812 提升至微調後的 0.820，其精確率、召回率與 F1 分數亦小幅提升至約 0.82。此結果顯示，進行針對特定領域的微調，能有效地提升語言模型對於財經文本中專業術語與情感線索的語境理解能力。

再次強調，由於本次研究在 Gemini 免費版微調樣本數有所限制，若能提升模型訓練樣本數與調整更細緻的參數設定，預期準確率等各項效能指標應能獲得進一步提升。

4.3 預測時間與運算成本分析

在評估各模型於實際財經情感分析應用中的可行性時，預測所需時間與運算成本是關鍵的考量指標。透過計算平均預測時間以及處理 516 筆句子的總耗時，我們可以全面評估各模型的執行效率。各模型在預測過程中所記錄的時間數據彙整如表 7 所示。

模型名稱	平均預測 (秒)	總預測 (秒)	總訓練 (秒)	運算處理器
`ft:gemini1.5flash`	2.353	2349	3425	A100
`base:gemini1.5flash`	1.498	1845	N/A	A100
`ft:finbert`	0.010	6.05	19260	A100
`ft:bert`	0.010	5.94	21360	A100
`t:svm`	0.002	0.92	479	CPU
`t:random-forest`	0.070	36.12	458	CPU
`t:logistic-regression`	0.0013	0.67	273	CPU

表 7: 各模型預測時間與成本比較

註：基礎版 Gemini 的訓練時間標為 N/A，因為它是零樣本預測。微調後 Gemini 的訓練時間 3425 秒是指在本次研究中針對少量樣本進行微調的時間。

Gemini 1.5 Flash 模型在零樣本條件且使用時間受限的情況下，成功預測 495 筆句子，平均每句預測時間為 1.498 秒，總處理時間為 1845 秒。經過微調的 Gemini 1.5 Flash 模型(`ft:gemini1.5flash`)成功預測 481 筆句子，平均每句預測時間略增至 2.353 秒，總處理時間為 2349 秒。若加上微調模型的訓練時間，總操作時間為 3425 秒。此結果顯示，儘管微調可以提升模型的效能指標，但同時也會增加運算時間，推測這與訓練過程中引入的額外參數與最佳化設定有關。

微調後的 `BERT` 與 `FinBERT` 模型在預測時間上明顯優於 Gemini 系列模型。它們的總訓練時間分別為 356 分鐘與 321 分鐘。在預測方面，兩者平均每句預測時間僅為 0.01 秒，總處理時間分別約為 7 秒與 6 秒。然而，儘管具備極高的運算效率，`BERT` 模型的預測表現在本研究中仍低於使用微調後 Gemini 1.5 Flash 的模型；而 `FinBERT` 模型的預測表現則優於微調後的 Gemini 1.5 Flash 模型。

羅吉斯迴歸是預測速度最快的模型，平均每句僅需 0.0013 秒。其次為 SVM（0.002 秒）與隨機森林（0.07 秒）。與 BERT 系列模型類似，是運算效率最高的選項。然而，傳統模型在準確率、精確率、召回率與 F1 分數等指標方面，皆明顯低於其他先進模型，顯示其高速表現是以犧牲情感分類效能為代價。

整體而言，結果顯示 FinBERT 模型雖然具有最高的準確率，但在選擇微調參數時，模型伴隨著較高的運算時間。而微調後的 Gemini 1.5 Flash 則提供一個相對高效的解決方案，惟其預測表現在本研究中仍不及微調後的 FinBERT 模型。傳統機器學習模型是最輕量化的選項，但在對準確率要求較高的情感分類任務中，其實用性有限。最終模型的選擇，應根據特定的財經應用場景，在準確性、效率與成本之間進行權衡判斷。

5 討論

在前述各章節中，我們針對大型語言模型（LLMs）、BERT 與 FinBERT 等自然語言處理模型，以及 SVM、隨機森林與羅吉斯迴歸等傳統機器學習模型，探討它們在財經情感分析與分類任務中的表現。首先，在零樣本（zero-shot）架構下評估基礎的 GPT 模型（此處指 Gemini），讓其在未經特定任務微調的情況下，直接判斷測試集中句子的情感屬性（正面、負面或中性）。接著，透過少樣本（few-shot）學習的方式對各模型進行微調，並在訓練集與驗證集上進行最佳化。對於 NLP 模型與傳統模型，則透過貝氏最佳化方式進行超參數調整，經過 100 次試驗後選出最佳的參數組合。

完成微調後，各模型對相同的測試集進行情感預測，並針對其表現進行詳細分析。本章節將綜合這些分類結果，進一步歸納研究發現並提出關鍵的洞見。

5.1 情感分類混淆矩陣分析

為更深入地瞭解各模型的分類結果，本研究透過視覺化的方式，呈現不同情感類別下各模型的優劣勢，並產製混淆矩陣熱力圖。熱力圖作為一種強大的視覺分析工具，能以直觀的方式呈現模型的分類效能，協助研究者掌握數據分布、特徵關聯與分類結果等複雜模式（S. Zhao 等人，2014）。在分類任務中，熱力圖常應用於混淆矩陣的呈現、特徵重要性分析以及深度學習模型激活映射的視覺化，有助於辨識錯誤分類、特徵貢獻與決策邊界。

在本研究中，混淆矩陣熱力圖中的對角線表示分類正確的情況，即模型預測的情感類別與實際標註的情感類別一致。就混淆矩陣而言，對角線即為各情感類別的「真陽性」（True Positives），是評估分類精準度的重要依據。

觀察基礎版與微調後的 GPT 模型結果（圖 1），可以發現它們在辨識中性情緒方面普遍存在困難，該類別的誤判率明顯偏高。例如微調前的 Gemini 模型，將 21 筆中性

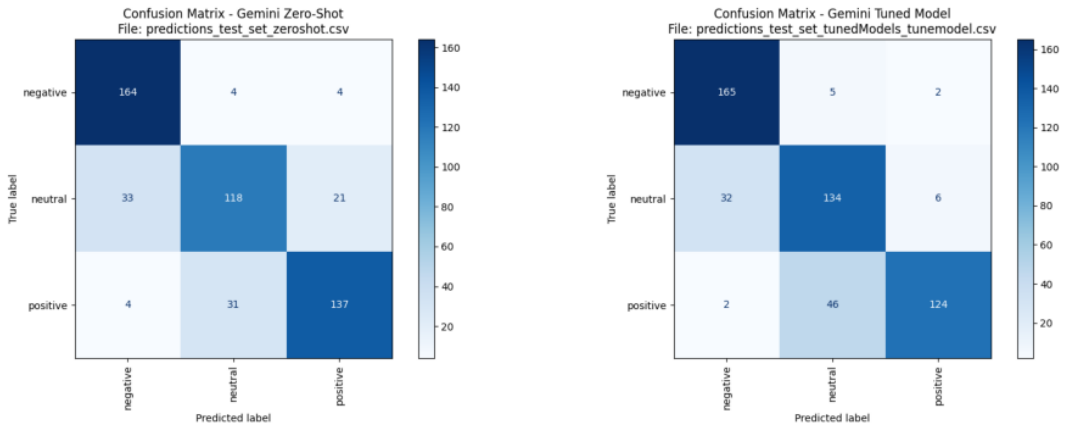

圖 1: GPT 模型在微調前後的預測結果與實際類別比較熱圖

樣本誤判為正面，這顯示模型在面對語意模糊或語氣平衡的文本時，可能傾向於以樂觀的角度解讀，而非保持中立。值得注意的是，微調後的 Gemini 模型在此方面的表現反而更佳，僅將 6 筆中性樣本誤判為正面。另一方面，微調前的 Gemini 模型，將 21 筆正性樣本誤判為中性，但經過微調後反倒將 46 筆正向樣本誤判為中性，這顯示其在某些語意細膩的情境中，可能具備更好的泛化能力。

如先前章節所述，金融文本中的中性情緒因其常具有隱晦且依賴語境的特性，在判別上具有高度挑戰性。表面看似中性的陳述可能隱含正面或負向的語意，若缺乏深入的語境理解，便極易被忽略。此一複雜性可以解釋為何無論是基礎版或微調後的模型，皆在中性樣本的分類上出現較高的錯誤率。此結果再次強調，在處理細緻的金融情緒時，需導入更進階的建模方法或更精緻的標註策略。

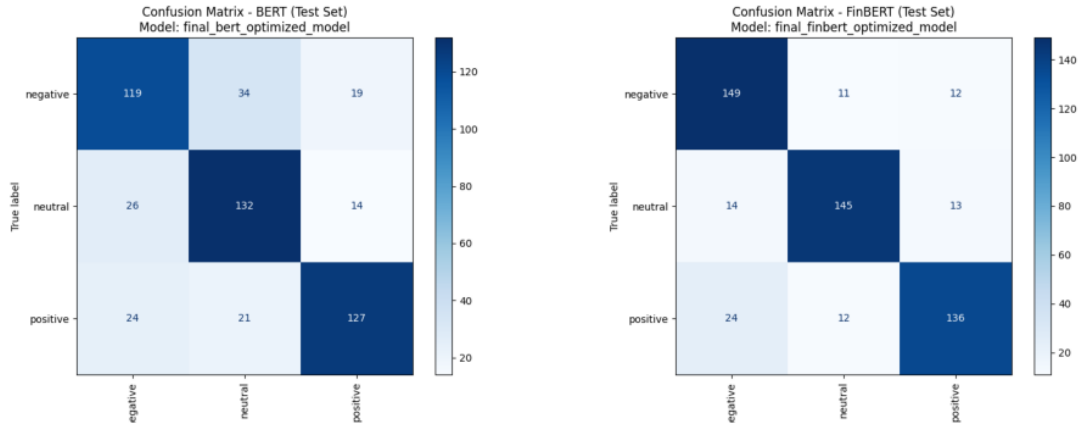

圖 2: BERT 與 FinBERT 模型預測結果與實際類別比較熱圖

圖 2 呈現兩種基於 BERT 的微調模型（BERT 與 FinBERT）之混淆矩陣熱力圖。從微調後的 BERT 模型開始觀察，其整體分類表現穩定，共正確預測 119 筆負面、127 筆正面及 132 筆中立樣本。然而，其主要的錯誤來源在於對中立情緒的誤判。具體而言，

BERT 將 34 筆負面樣本誤判為中立，這顯示其可能對金融文本中常見的保守或謹慎語言較為敏感。此外，有 26 筆中立樣本被預測為負面，另有 24 筆正面樣本被誤判為負面。雖然這些錯誤的數量有限，但仍反映出模型在辨識溫和與高度極化情緒時，對細微語言線索的處理仍有挑戰。

在相同資料集上進行微調的 FinBERT 模型，在處理語意模糊的情境下表現略優。作為一個專為金融領域訓練的 Transformer 模型，FinBERT 能正確分類 149 筆負面、136 筆正面與 136 筆中立樣本。與 BERT 相比，FinBERT 在辨識負面情感時的錯誤較少，僅將 11 筆負面樣本誤判為中立，12 筆負面樣本誤判為正面。有趣的是，也有 24 筆正面樣本被預測為負面，這顯示 FinBERT 偶爾會低估明顯正面語句的情感極性。然而，其在中立情感處理上的改進，顯示針對特定領域的預訓練有助於降低在高度依賴語境的分類任務中的混淆程度。

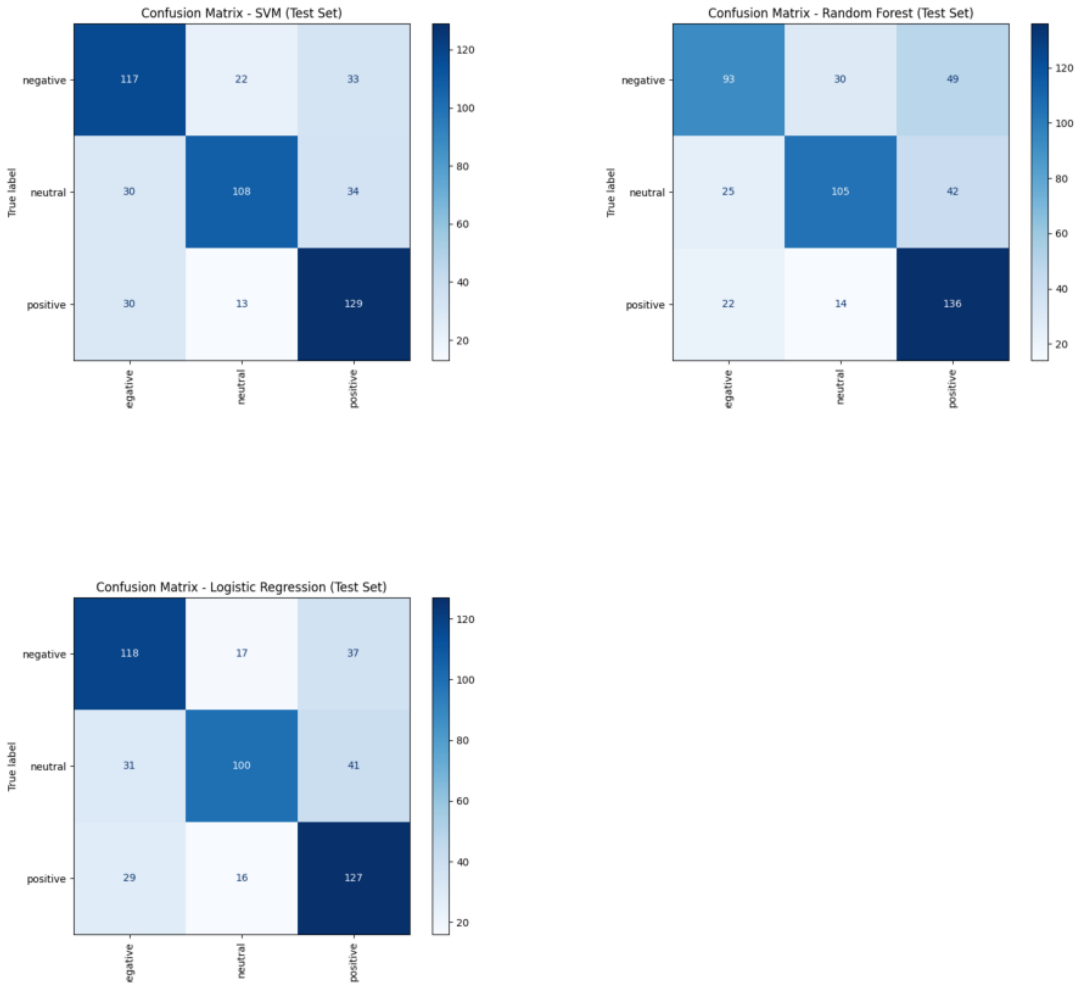

圖 3: 傳統機器學習模型預測結果與實際類別比較熱圖

圖 3 呈現三種傳統機器學習微調模型（SVM、隨機森林與羅吉斯迴歸）的混淆矩陣熱力圖。首先觀察 SVM 模型，其在情感區分上的表現明顯下降。SVM 僅正確預測 117 筆負面與 108 筆中立樣本，而正面情感的分類相對較佳，正確分類 129 筆。不過，模型在情感類別間出現大量的混淆：有 34 筆中立樣本被標記為正面，30 筆中立樣本也被誤判為負面。此外，另有 33 筆負面樣本被預測為正面，30 筆正面樣本則被誤判為負面。這些結果顯示 SVM 難以處理較為複雜且非線性的語意關係，特別是在面對語氣平衡或間接表述的情境下，其情感辨識能力顯得不足。

羅吉斯迴歸相較於 SVM 表現略有下降，分別正確分類 118 筆負面、127 筆正面與 100 筆中立樣本。可發現模型在正面與其他情感類別之間的混淆仍然明顯。例如，有 41 筆中立樣本被誤判為正面，37 筆負面樣本亦被誤判為正面。同樣地，另有 31 筆中立樣本則被分類為負面。這顯示即使採用提升模型穩健性的方法，模型仍缺乏足夠的語意理解能力，難以正確解析金融語句中隱含的情緒線索，特別是在情感極性並非明確表達的邊界案例中。

最後，隨機森林分類器在情緒分類任務中表現最為不足。其雖正確預測 93 筆負向、136 筆正向與 105 筆中性樣本，但在負向與中立類別上出現大量誤判。具體而言，有 49 筆負向與 42 筆中性樣本被誤判為正向；另有 30 筆負面樣本則被分類為中立。此結果顯示，當模型面對語意不明確的句子時，傾向預測為正向，反映其在處理複雜句構或隱含情緒表達方面的能力有限。

總結而言，研究結果顯示各模型在分類中性情緒時普遍面臨困難。此挑戰在未經財經領域特化訓練的模型中尤為明顯，因為財經文本常見的模糊表述或保留語氣，易使中性與輕微正負向情緒的界線模糊。FinBERT 的表現凸顯領域專屬預訓練的價值，而 SVM 等傳統機器學習模型則反映其在掌握語境細節與隱含情緒方面的侷限性。

5.2 研究發現與限制

第四章的實驗結果明確顯示，基於 Transformer 的模型相較於傳統分類器具有顯著優勢。具體而言，BERT 與 FinBERT 等 Transformer 模型的平均準確率較傳統模型（SVM、隨機森林、羅吉斯迴歸）高出 11.18%。儘管傳統模型的平均準確率僅為 67.12%，但仍應將此結果置於適當的脈絡下解讀。考量到隨機分類的正確率僅為三分之一（約 33%），傳統模型已顯著優於隨機基準，這顯示其雖不及 Transformer 模型，但仍具有一定的應用價值。

Transformer 模型從根本上重塑人工智慧的發展，尤其是在自然語言處理領域。BERT 模型已廣泛應用於各類分類與 NLP 任務中，並持續展現其效能與成本效益。然而，生成式 AI 的出現進一步推動 NLP 領域的進展，大型語言模型（LLMs）在效率上已超越或至少可與傳統 NLP 模型相當。這種效率的提升主要來自於其在數十億參數規

模上的大規模預訓練，使其能廣泛泛化於各種任務。值得注意的是，LLMs 的真正優勢不僅在於預訓練的規模，更在於其具備微調能力，使其能有效地適應原始訓練範疇以外的任務。

在可用的訓練資料有限的情況下，經過微調後的 LLMs 表現仍無法超越 FinBERT，但相較於其原始的基礎大型語言模型，平均準確率提高 0.96%，可預期未來增加訓練樣本與模型複雜度後可再提升準確率。在訓練時間方面，FinBERT 微調需時 19,260 秒，而 Gemini 在有限資料上的微調所需時間則大幅減少約 82.77%。若將 LLM 與傳統分類器比較，訓練時間的差異更加顯著。然而，在預測階段觀察到不同趨勢：LLMs 的預測成本與時間均高於 BERT 與傳統模型，增加約 150%，因此對於即時應用而言更需審慎評估其效益與可行性。

儘管存在上述挑戰，LLMs 仍具高度潛力，特別是在採用參數效率高的微調技術後更具實用性。本研究中，BERT 與傳統模型均透過貝氏最佳化進行完整的超參數調整，以找出最適配置；然而，LLMs 的微調則使用預設的超參數。超參數最佳化對於重視效能的實務應用尤其關鍵，但在 LLMs 中，每組超參數組合皆需獨立執行微調作業，成本極高。此外，傳統機器學習中常用以提升效率的提早停止（early stopping）技術，於封閉式 API 架構的 LLMs 中通常不可用，進一步限制其調校效率。

根據本研究結果，轉換器模型的優異表現與適應性，對金融科技產業具有直接的應用價值。像 BERT 與經微調後的 LLM，可應用於金融平台中，自動分析新聞標題、法規更新與財報公告的情緒，提供即時洞察，輔助風險評估與投資決策。此外，投資者儀表板亦可結合這些模型，用以標示市場訊號、偵測情緒變化，或自動生成摘要報告，提升機構與散戶投資決策流程的效率。由於 LLM 成本較高，因此模型的選擇將取決於效能與資源間的權衡：規模較小的平台可能偏好成本效益佳的 BERT 模型，而資源充足的大型機構則可受惠於 LLM 較高的準確性。

6 結論

本研究結果突顯金融領域情感分類技術的發展趨勢，各類模型皆展現出不同的優勢與需要權衡的面向。Transformer 類型的模型，特別是 FinBERT，在處理中性語句方面表現優異，對於中立性具有重要意涵的財經文本尤其具有應用價值。相較之下，傳統分類器雖然在準確率上不如 Transformer 模型，但在訓練資料量較大時仍具可行性。其計算成本低、訓練與預測速度快的特性，亦使其適用於預算受限下的即時情境。

另一方面，大型語言模型（LLMs），特別是經過微調的模型（如本研究中使用的 Gemini），在準確率方面始終維持相當高的表現，明顯優於除 FinBERT 以外的其他模型。然而，此類表現需付出高昂的代價，包括極高的運算成本，以及多項限制，如

API 使用限制、缺乏透明度與難以應用高效率的訓練策略等問題。

綜合上述結果，可根據應用情境採取分層式的模型選擇策略，像是在成本敏感且需處理大量資料的任務中，傳統模型為合適的選擇。若需兼顧效能與可解釋性，則可採用如 `BERT` 與 `FinBERT` 等 Transformer 模型。而在資料有限且準確性極為關鍵的高風險場景中，則建議採用大型語言模型（LLMs）。

未來研究應聚焦於參數效率高的微調技術與具成本意識的最佳化策略，以縮小效能與效率間的落差，促進 LLMs 在金融實務應用中的廣泛落地。本研究的洞見指出，儘管像 `FinBERT` 和微調後的 LLMs 這樣的高階模型能提供更優越的準確度，但最佳工具的選擇最終仍取決於特定金融情感分析任務的具體限制與目標。

參考文獻

[1] Du, K., Xing, F., Mao, R. & Cambria, E. (2024). Financial Sentiment analysis: Techniques and applications. *ACM Computing Surveys, 56*(9), 220.

[2] Du, K., Xing, F. & Cambria, E. (2023). Incorporating multiple knowledge sources for targeted aspect-based financial sentiment analysis. *ACM Transactions on Management Information Systems, 14*(3), 23.

[3] Malo, P., Sinha, A., Korhonen, P., Wallenius, J. & Takala, P. (2014). Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology, 65*(4), 782–796.

[4] Liu, Z., Huang, D., Huang, K., Li, Z. & Zhao, J. *FinBERT: A pre-trained financial language representation model for financial text mining.*

[5] Yang, Y., Christopher, M., Uy, S. & Huang, A. (2020b). *FinBERT: A pretrained language model for financial communications.*

[6] Yang, Y., Christopher, M., UY, S. & Huang, A. (2020a). *GitHub—yya518/FinBERT: A pretrained BERT model for financial communications.* https://github.com/yya518/FinBERT

[7] Dmonte, A., Ko, E., & Zampieri, M. (2024, December 15–18). An evaluation of large language models in financial sentiment analysis. In *2024 IEEE International Conference on Big Data (BigData)* (pp. 4869–4874). Washington DC, USA: IEEE.

[8] Fatouros, G., Soldatos, J., Kouroumali, K., Makridis, G., & Kyriazis, D. (2023). Transforming sentiment analysis in the financial domain with ChatGPT. *Machine Learning with Applications, 14*, 100508.

[9] Malo, P. & Sinha, A. (2024). Financial PhraseBank. https://huggingface.co/datasets/takala/financial_phrasebank

[10] Sbhatti. (2021). *Financial sentiment analysis*. https://www.kaggle.com/datasets/sbhatti/financial-sentiment-analysis (存取日期：2025/4/20).

[11] Hugging Face. (2024). BERT base uncased model. https://huggingface.co/bert-base-uncased

[12] Zhang, K., Zhou, F., Wu, L., Xie, N., & He, Z. (2024). Semantic understanding and prompt engineering for large-scale traffic data imputation. *Information Fusion, 102*, 102038.

[13] Zhao, S., Guo, Y., Sheng, Q., & Shyr, Y. (2014). Advanced heat map and clustering analysis using Heatmap3. *BioMed Research International, 2014*(1), 986048.