**語意分析和長短期記憶用於新聞預測股市未來漲跌走勢**

Using Semantic Analysis and LSTM to Predict the Stock Market Trend with Financial News

**指導教授：陳牧言Mu-Yen Chen**

**研究生：謝仁堡Ren-Pao Hsieh**

**國立臺中科技大學**

**資訊管理系碩士班**

**碩士論文**

A Thesis

Submitted to

Department of Information Management

National Taichung University of Science and Technology

in Partial Fulfillment of the Requirements

for the Degree of

Master of Information Management

**July 2018**

**Taichung, Taiwan, Republic of China**

**中華民國107年7月**

Using Semantic Analysis and LSTM to Predict the Stock Market Trend with Financial News

|  |  |
| --- | --- |
| Student: Ren-Pao Hsieh | Advisor: Mu-Yen Chen |

Department of Information Management

National Taichung University of Science and Technology

## ABSTRACT

The science and technology is more and more developed. Digital media such as articles, commentary, videos, animations and others on the Internet is becoming more and more important. English semantic analysis has many basic technologies, many applications are also gradually budding in this basic technology. On the other hand, there is no uniform or complete reorganization of the basic technologies in Chinese semantic analysis. Chinese semantic analysis is difficult than English semantic analysis because it is difficult to judge the true meaning of Chinese words and sentences. This study collects articles about common news sites in Taiwan and related to individual stocks. After the data is preprocessed and Skip-gram, each word is converted to word features using Word2Vec. The Lexicon stores the most relevant words around the keyword. In the prediction stage, this study calculates the impact of new articles on the stock price according to the full training lexicon. Finally, this study uses the LSTM to evaluate the final results. This study predicts future stock market trend through data processing and analysis through news articles.

Keyword：*Semantic analysis, stock market trend analysis, time series, Long Short-Term Memory (LSTM)*

語意分析和長短期記憶用於新聞預測股市未來漲跌走勢

|  |  |
| --- | --- |
| 學生：謝仁堡 | 指導教授：陳牧言 |

國立臺中科技大學資訊管理系碩士班

## 摘要

隨著網路越來越發達，更注重網路上的文章、評論甚至是影片、動畫等等數位媒體，因此這些數位媒體所產生的數據變得越來越重要，在語意分析中，英文文本分析的發展較為健全，許多基礎技術已經很完善，應用面也相對的容易許多，相較之下，中文文本分析具備的基礎技術並沒有統一或完整的整頓，皆為零散的片段技術，不易根據單詞和句型判斷真正意涵，中文語意分析難度也隨之上升。本研究從台灣常見的新聞網站蒐集金融相關的文章，採用語意分析、文本探勘分析新聞文章與股市價格的關係並預測未來股價走勢，其資料經過預處理、斷字後經過Word2Vec轉化向量作為每一個詞彙特徵，詞彙特徵會篩選關鍵字周圍最相關的詞作為建立詞庫的依據，預測階段根據完整訓練後的詞庫對應新文章內容，並判定新文章對於股價的影響。最後，我們使用LSTM評估結果，本實驗中透過新聞文章的分析和資料愈處理預測股市趨勢。

關鍵詞：*語意分析、股市預測、時間序列、長短期記憶、LSTM*

## 目錄

[ABSTRACT I](#_Toc519447627)

[摘要 II](#_Toc519447628)

[目錄 III](#_Toc519447629)

[圖目錄 IV](#_Toc519447630)

[表目錄 V](#_Toc519447631)

[壹、 前言 1](#_Toc519447632)

[一、 背景與動機 1](#_Toc519447634)

[二、 目的 5](#_Toc519447635)

[貳、 文獻探討 6](#_Toc519447638)

[一、 語意分析 6](#_Toc519447639)

[二、 循環神經網路(Recurrent Neural Network, RNN) 8](#_Toc519447641)

[三、 長短期記憶(Long Short-Term Memory, LSTM) 10](#_Toc519447642)

[四、 斷詞 11](#_Toc519447643)

[五、 詞向量(Word2Vec) 12](#_Toc519447644)

[六、 關鍵詞判定 15](#_Toc519447645)

[七、 股市應用相關研究 18](#_Toc519447646)

[參、 實驗流程 20](#_Toc519447647)

[一、 資料描述 20](#_Toc519447649)

[二、 實驗流程 21](#_Toc519447650)

[肆、 實驗設計與結果 33](#_Toc519447651)

[一、 個股(三類股)台積電、鴻海、台塑化 33](#_Toc519447653)

[二、 台灣50預測大盤指數 39](#_Toc519447654)

[三、 台灣大學公開的極性詞庫(NTUSD) 42](#_Toc519447655)

[四、 實驗結果討論 49](#_Toc519447656)

[伍、 結論 51](#_Toc519447657)

[一、 研究貢獻 51](#_Toc519447658)

[二、 研究限制 52](#_Toc519447659)

[三、 未來研究方向 53](#_Toc519447660)

[參考文獻 54](#_Toc519447661)

## **圖目錄**

[圖 1.1 全球物聯網支出 2](#_Toc518509784)

[圖 2.1循環神經網路概念圖 9](#_Toc518509785)

[圖 2.2 LSTM結構圖 10](#_Toc518509786)

[圖 2.3 字詞執行數值編碼 12](#_Toc518509787)

[圖 2.4 Word2vec類神經網路擷取特徵向量，單詞以四個維度表示為例 13](#_Toc518509788)

[圖 2.5 Word2Vec類神經網路句子模型 14](#_Toc518509789)

[圖 2.6 詞向量訓練架構 15](#_Toc518509790)

[圖 3.1 研究流程圖 20](#_Toc518509791)

[圖 3.2二維空間平面表示斷字後的向量特徵以2016/01/18，以Lag-1為例 24](#_Toc518509792)

[圖 3.3 以二維空間中表示關鍵字周圍的100字以鴻海新聞2016/01/18，以Lag-1為例 25](#_Toc518509793)

[圖 3.4 LSTM架構圖 30](#_Toc518509794)

[圖 3.5 LSTM Dropout位置圖 31](#_Toc518509795)

[圖 3.6 2維資料攤平成1維資料 32](#_Toc518509796)

[圖 4.1 本研究詞庫分析個股、大盤股價對於新聞的Lag 41](#_Toc518509797)

[圖 4.2 本研究與台灣大學詞庫分析個股新聞文章最佳結果比較 49](#_Toc518509798)

## **表目錄**

[表 2.1 詞類標籤比較表 22](#_Toc519093756)

[表 2.2 詞性類別定義與細項分類 23](#_Toc519093757)

[表 2.3 字詞統計 26](#_Toc519093758)

[表 2.4 字詞正、負向的權重總和 26](#_Toc519093759)

[表 2.5 平均權重 27](#_Toc519093760)

[表 2.6鴻海(2317) 每日數據與正規化 28](#_Toc519093761)

[表 2.7 每日正、負文章量，以鴻海為例 29](#_Toc519093762)

[表 2.8 LSTM輸入數據，以鴻海Lag-1為例 31](#_Toc519093763)

[表 3.1使用本研究詞庫分析台積電新聞所使用的參數 34](#_Toc519093764)

[表 3.2使用本研究詞庫和不同Lag分析台積電新聞 34](#_Toc519093765)

[表 3.3使用本研究詞庫和不同迭代次數分析台積電新聞 34](#_Toc519093766)

[表 3.4使用本研究詞庫分析鴻海新聞所使用的參數 35](#_Toc519093767)

[表 3.5使用本研究詞庫和不同的Lag分析鴻海新聞 35](#_Toc519093768)

[表 3.6使用本研究詞庫和不同迭代次數分析鴻海新聞 36](#_Toc519093769)

[表 3.7使用本研究詞庫分析台塑化新聞所使用的參數 36](#_Toc519093770)

[表 3.8使用本研究詞庫和不同的Lag分析台塑化新聞 37](#_Toc519093771)

[表 3.9使用本研究詞庫和不同迭代次數分析台塑化新聞 37](#_Toc519093772)

[表 3.10使用本研究詞庫分析個股新聞相同迭代次數訓練時間 38](#_Toc519093773)

[表 3.11使用本研究詞庫分析個股新聞最佳結果 39](#_Toc519093774)

[表 3.12使用本研究詞庫分析台灣50相關新聞所使用的參數 40](#_Toc519093775)

[表 3.13使用本研究詞庫和不同的Lag分析台灣50相關新聞 40](#_Toc519093776)

[表 3.14使用本研究詞庫和不同迭代次數分析台灣50相關新聞 40](#_Toc519093777)

[表 3.15使用本研究詞庫分析個股與台灣50相關新聞最佳結果 41](#_Toc519093778)

[表 3.16 NTUSD詞庫分析台積電新聞所使用的參數 42](#_Toc519093779)

[表 3.17 NTUSD詞庫和不同的Lag分析台積電新聞 43](#_Toc519093780)

[表 3.18 NTUSD詞庫和不同迭代次數分析台積電新聞 43](#_Toc519093781)

[表 3.19 NTUSD詞庫分析鴻海新聞所使用的參數 44](#_Toc519093782)

[表 3.20 NTUSD詞庫和不同的Lag分析鴻海新聞 44](#_Toc519093783)

[表 3.21 NTUSD詞庫和不同迭代次數分析鴻海新聞 44](#_Toc519093784)

[表 3.22 NTUSD詞庫分析台塑化新聞所使用的參數 45](#_Toc519093785)

[表 3.23 NTUSD詞庫和不同的Lag分析台塑化新聞 45](#_Toc519093786)

[表 3.24 NTUSD詞庫和不同迭代次數分析台塑化新聞 46](#_Toc519093787)

[表 3.25 NTUSD詞庫分析台灣50相關新聞所使用的參數 47](#_Toc519093788)

[表 3.26 NTUSD詞庫和不同的Lag分析台灣50相關新聞 47](#_Toc519093789)

[表 3.27 NTUSD詞庫和不同迭代次數分析台灣50相關新聞 47](#_Toc519093790)

[表 3.28 NTUSD詞庫分析個股與大盤最佳結果 48](#_Toc519093791)

[表 3.29使用本研究詞庫與NTUSD詞庫最佳結果比較表 50](#_Toc519093792)

## 前言

# 

### 背景與動機

現階段雲端、物聯網等技術的結合，數據量以指數成長，根據IDC 2017年的研究報告指出，2013年至2020年資料將成長10倍，總量將從4.4ZB增加至44ZB，其中三分之二的資料是由個人產生，物聯網在全球的花費上逐年上升，如圖 1.1所示，可想而知個人數據越來越重要。對於論壇網站像是Twitter、PTT、Facebook、微博等等，將是未來大量數據的提供者，在這些網站中都是以自由言論為基礎，許多數據是屬於非結構的文本，Pandey (2017)表示現今論壇網站已經改變人們在線交流的格局，通過網路創造專屬於自己空間的內容，Meire (2016)提到情緒分析主要是在審查數據上進行的。由於社交媒體蓬勃發展，已成為該領域的主要焦點，但沒有固定的方式能夠自動解密這些文本，必須依據動態、人工的方式分析，才能從文本資料解析成有價值的資訊。Hung (2017)表示口碑是可以有效影響消費者購買決策的重要信息，而消費者可以輕易地找到其他消費者的經驗或評論作為參考。由此可知，論壇、社交網站中的評論變成消費者重視的口碑之一，線上互動的數據是屬於寬鬆、無限制的格式文本，提升擷取關鍵字詞的難度，Giatsoglou (2017)採用情緒分析和意見挖掘從文本文檔中提取有用的主觀信息是有價值的。良好的預處理能夠過濾雜訊，能夠減緩處理的時間、提升訓練效果，相較過去的數據量，在限制時間內找出有用的特徵，花費的成本將會提升許多。

圖 1.1 全球物聯網支出

以往的資料分析將固定格式的新聞、期刊論文、文章，關鍵字搜尋已被廣泛運用，能夠將上述資料，根據文章結構擷取特徵，定義關鍵字眼、找出文章重點，過往的分析技術僅能針對有明顯結構的文章分析，針對非結構文本，像是散文、使用者發文、小說等，是有一定的難度，甚至有些是圖文文本，又是一個進階的挑戰。Fu (2017)提到傳統的文本分析方式是採用複雜的人工擷取特徵，過程中耗費大量時間與成本，在建立足夠的詞庫的情況下，能夠將傳統耗時分析的缺點改善，並且得到更好的結果。

金融市場從以往的現場資訊慢慢轉換為網路資訊，交易模式和獲取方式也隨之改變，許多投資者在網路上獲取個股相關資訊，這些資訊無時無刻都在變動，掌握最新資訊，轉換自身的投資策略，最後獲得利益。相關資訊像是成交股數、金額、交易筆數、最高最低價、開盤收盤價、外資交易數據，還有一些廣泛運用的曲線和技術分析像是日均線、月均線、KD指標、RSI指標、MACD指標、乖離率BIAS，每個人不同的想法、不同的運用方式，當然投資策略也略有不同，這些技術分析都是根據獲得的數據經過運算後所呈現，觀看技術分析曲線選擇適合自己的股票，便是這些分析最主要的價值所在，除了這些一般能夠獲得的資訊外，許多投資人都會觀看個股新聞和其評論作為投資策略之一。

電子商務的環境下，許多消費者將其感受回饋給公司、廠商，而其他消費者透過評論的分享，進而瞭解產品的真實性，增進消費者對於該產品的信任並增加購買意願，然而有些企業發現消費者相當重視評論，聘僱寫手操弄評論提升銷售額的手法越來越常見，大量訊息中僅能依靠經驗辨識其評論真偽。若將其運用於金融新聞，評論一面倒向負面，可想而知其他投資者已經透露出該股未來走向相當不樂觀，也許是有心人士撰寫，但投資客並不會查證每一句、每一篇評論的真實性，且投資客並不會將所有評論完整看完，只能觀看大量的評論後，擷取片段評論，當然也很難證實每一篇評論中沒有擷取含有操弄的評論，相對的就無法察覺哪些才是真實評論。

現今有許多因素影響股市，不僅僅是數據、技術分析、評論，還有投資者情緒會影響投資決策，可能會透過網路迅速傳播擴大，在股市上可能會受到一定程度的影響，Guo (2017)認為情緒較為悲觀時，觀看股市資訊和制訂決策時往往會偏向個人情緒。有些研究探討Twitter、Google中提取的新聞和民意趨勢和財經新聞等等，García (2016)表示社交媒體對於股市變化和預測跡象，各種社交媒體中表示的是真實的民眾觀點，而這些民眾可能對於股市交易具有一定的影響力，投資客除了觀看社群媒體上的評論、文字外，還有影片、音訊等資訊也是能夠判定的數據，且動態數據越來越受到矚目，甚至許多廣告是透過周圍環境的改變而產生變化。在數位學習上也是一個指標，再也不是傳統文本文字的方式教學，身歷其境的感受最容易讓使用者、受教者感到有趣且容易記憶，觀察到這點的研究者，將影音也納入影響投資客的作判斷決策的要素。

目前面臨數據的快速增長，財務文本也不例外，而這些大量的數據對於許多組織來說是一個很大的挑戰，分析工具的需求也是不斷的增進與改變，一般而言，文本比數字數據處理更具挑戰性，Chan (2017)認為文本是屬於非結構化的資料，在財務決策中也是個有價值的資料集。當前數據電子化的程度已經越來越高，許多傳統文件也透過電腦設備轉化為電子檔儲存備份，甚至透過網路上傳分享數據，分析人員該如何應用這些數據將會是企業未來價值的關鍵。

語意分析、文本探勘近年來廣泛運用，英文文本分析的發展較為健全，許多基礎技術已經很完善，應用也相對容易許多，中文文本分析具備的基礎技術並沒有統一或完整的整頓，皆為零散的片段技術，在發展上，中文語意分析已落後英文一段時間，再加上英文詞彙具有明顯結構格式，可輕易的建立詞彙關聯模型。Wang (2016)主張中文語意分析與英文語意分析具有顯著差異，若將情緒分析應用於英文的環境下衡量中文語意，因中英文化或語言結構差異，無法衡量真正語意。Poria (2017)表示漢字的性質，無法將英文的研究直接使用於中文研究。慶幸的是能夠將其概念試著引入中文，像是語法、詞性建置，一個詞是由多個字所組成，每個詞都具有意義，該如何了解表達者真正想表達的意涵，如何擷取真正重要的單詞特徵成為中文語意分析一大挑戰。目前中文句型結構分析還有一定的難度，英文擁有一定的格式與斷字，較容易擷取文字，中文詞彙的淺在意義或隱含意義當前沒有定義統一標準，嘲諷語氣、反向語氣許多語意分析的文獻甚至沒有特別說明，當然定義上會有許多問題，現階段只能採用相對之下較合理的方式執行研究。對於傳統的語意分析，從資料收集到詞庫製作完成可分為5個步驟，由圖 1.2做介紹，其中第四步驟為關鍵步驟，且需要找到該領域專家認定，透過人工認定方式相當主觀，較符合人類觀點訓練的模型更容易擬和人類真實想法，但花費時間最長，且專家認定的方式每次擴充必須再次執行。



圖 1.2語意分析人工認定

### 目的

當前數據快速成長，每個人能記憶的數據量都是有限的，求解、搜尋的過程中該如何過濾不必要的雜訊是一門很重要的學問，光是篩選除了要具備專業能力、相關知識外，最主要的是要花上大量時間，節省時間成本也成了當前主要的考量因素。為了深入了解文句中所表示的意涵、標籤極性，文本分析處理數據相當費時，字詞定義無法客觀定義，專家認定又過於繁瑣，資料收集針對擁有目標欄位的資料作分析較有效益。傳統語意分析認定流程步驟需要5個步驟，本研究定義詞庫只需要3個步驟，能夠有效減少實驗流程複雜性，認定方式中，傳統語意分析透過人工方式認定，多少擁有主觀意見與想法，本研究採用統計的方式，讓模型自行透過訓練文本與目標欄位之間產生對應的極性，相較之下較為客觀且擴充性較高。許多文獻探討一篇文章該如何找到文章關鍵字與特徵，其中定義描述會有很大的問題，本研究探討股市影響因素，投資者觀看數據、新聞制定決策，根據自身過往經驗預測未來走勢，在大數據和歷史資料的情況之下證明這些數據對於未來是有影響力的，透過長短期記憶(Long Short-Term Memory, LSTM)，將特徵合併訓練過往資料，使模型學習目標欄位具備的特徵與權重參數的更新調整，根據實驗訓練的最佳模型預測個股未來股價走勢。

* 以統計權重優化傳統語意分析流程
* 財金新聞文章對於個股、大盤指數的影響
* 個股相關文章對於股價的
* 採用LSTM獲取歷史資料特徵

# 

# 

## 文獻探討

### 語意分析

# 

語意分析相關研究著重於網頁的資料探勘、文字探勘，這兩個議題表示如何從大量文字中和內容文章中挖掘出有用的價值資訊，傳統的語意分析、情緒探勘根據字詞預測文本表示的內容是屬於正向、負向、情感的探勘。Lin (2018)提到文本的情緒探勘研究工作可分為三個層次：文本層次、句子層次和實體層次，近期有許多研究朝向研究在文本數據中找出情感的詞彙，Guo (2017)表示投資者的情緒會影響他們的決策。透過網路的快速傳播和放大，股市上會受到一定程度的影響，將文章擷取特徵後，根據數據資料判定與訓練，準確的判定文章的目標情緒，其中包含許多演算法，監督式學習、半監督式學習、非監督式學習，分析的工具採用R、Python、Statica、Weka等工具都能夠做到資料分析，Deng (2017)針對監督式學習使用大型註釋詞庫訓練機器學習分類器，其中每個文本的標籤情緒被專家認可，專家標籤每個文本需要耗費大量的時間。根據逐一標記能夠建立具有真實人類情感的文章，使用機器學習，根據標記文本擷取特徵並與目標類別學習特徵，大量文本中若將所有詞彙一併訓練，可能會因為過多的雜訊導致分析與實際差異過大、收斂時間過長，使用一些特徵擷取的方式從文本中提取權重較高的詞彙是文本分析常見的做法，其中Chan (2017)也表示TF-IDF (Term Frequency-Inverse Document Frequency)衡量詞彙權重在大量數據中探索的應用中使廣泛常見的標準加權因子，搭配權重學習的詞彙。當中技術不論是統計的方法或是模型的方式計算，將訓練模型更精準的篩選出真正重要的詞彙作為訓練的依據，訓練完成後可依據此模型預測新文本特徵的目標類別。

文字探勘歸納於語意分析技術中，將文章、文本分句子段落、詞彙段落進行分析計算，取得詞彙之間的意涵與語意，或是句子整個架構中所代表的意義，Zhang (2015)提到近幾年的資訊量快速成長，研究人員也開始注意消費評論的重要性，該如何從中擷取關鍵詞彙，語意分類根據極性分類正負向為當前最主要的研究，語意分析能使用在評論分析，也能夠運用在其他非結構化文本中，像是Twitter、Facebook等內容中，根據使用者表達的文字和習慣、貼圖表情符號分析當前使用者發表這篇文章時的情緒，Ortigosa (2014)認為傳統的分析方式採用人工擷取特徵，過程中耗費大量時間與成本，若採用機器學習的方式自動判定將會節省許多。Araque (2017)在研究中衡量成本與成效後，機器判定能以最小的成本獲得最大的成效。因此在語意分析時，應採用機器判定的方式，經過大量資料的訓練，降低誤差值，不論是極性判定或詞語預測都能適用此概念，期望能夠達到專家判定的精準度，節省時間上的成本，並換取最高的效益。

到目前為止已有許多文獻探討語意分析，多半屬於英文語意而非中文，本章介紹過去文獻與研究方法，本研究蒐集了四篇期刊，分別為英文線上論壇焦點和預測、道瓊指數股市預測、使用頻率模式樹(Frequent Pattern-Tree, FP-Tree)演算法在多核心系統下平行處理文字探勘、Facebook動態牆情緒分析，以下依序介紹。

Li et al. (2010) 使用K-means、SVM(Support Vector Machine)預測英文線上論壇的焦點。此篇文獻為典型基礎應用，在某個主題、論壇中找尋熱門關鍵字，在線論壇熱點檢測和預測使用情緒分析和文本挖掘的方式，並創建一個算法來自動分析文本的情感極性並且獲得每個文本的值。結合這種算法用K-means和SVM開發無監督文本挖掘方法。我們使用提出的文本挖掘方法將論壇分組成各個集群，每個集群的中心代表當前時間範圍內的熱點論壇。

Bollena et al. (2011) 採用時間序列(Time series)、OpinionFinder、GPOMS（Google Profile of Mood States）預測道瓊指數(Dow Jones Industrial Average, DJIA)。對於中文分析來說相當困難進行相同任務分析，研究結果表示每日Twitter Feed的文本內容測量積極與消極情緒，以及測量情緒的Google心情狀態GPOMS在6維度（平靜，警告，確定，熱絡，親切和快樂）。由OpinionFinder測量的公眾情緒狀態的假設和GPOMS心情時間序列，都預示著DJIA收盤價值的變化。

Gadia et al. (2015) 主要根據FP-Tree、頻率模式成長樹(Frequent Pattern Growth, FP-Growth)兩個研究方法快速地找出需求文字。其中採用的研究方法，擁有較高的時間複雜度，由於此篇資料庫過於龐大，過去方法會造成過大的時間成本，因此在多核心機器上使用FP-Growth演算法，平行處理資料，規則建立完成後，將其運用和分析新聞文章，能有效改善以往方法的缺點。

Ortigosa et al. (2014) 採用決策樹(Decision Tree, DT)、簡易貝式(Naïve-Bayes)、SVM(Support Vector Machine)等方法，經過語意分析後，判斷使用者個人在論壇中發表的文章中，探討個人情緒變化帶來的學習影響。此篇文獻的前處理相當詳細，能夠給予相當的影響，但兩者之間具有明顯的語言差距，需要突破中文許多限制，在實作上具有一定的難度。

上述文獻中語意分析可分為目標欄位、已定義詞庫極性兩種方式，認定每一個詞所具有的極性，再根據每一個詞所賦予的標籤做分析，針對目標欄位認定字詞極性的方式，若沒有客觀的流程、方法定義目標欄位，與自行主觀認定詞庫的方式相同，有失研究客觀性，因此語意分析採詞庫認定的方式，多半採用現今已公開且可擷取的詞庫作為認定標準，各自研究並不會對於此多加描述，甚至不會去探討詞庫的內容，其中Bollena et al. (2011) 的研究架構，本研究將其作為文章認定參考依據，且同樣使用Time Seriese概念進行研究，主要差異在於後續分析評估，本研究將極性標籤後的文章數量作為LSTM的輸入數據，以目的的觀點探討與其餘文獻差異點，文獻中採用不同的演算法評估，比較結果、時間、挖掘熱門焦點，本研究探討新聞文章與股價關係，採用LSTM作為評估方式。

### 循環神經網路(Recurrent Neural Network, RNN)

本研究探討循環神經網路(Recurrent Neural Network)，而非遞迴神經網絡(Recursive Neural Network)，兩者之間有很大的不同，目前文獻探討的屬於循環神經網絡，至於遞迴神經網絡採用樹狀結構處理序列數據，耗費大量運算成本，在文本計算中已存在許多複雜的運算。

傳統類神經無法探討前一個序列與下一個序列之間的關係，在t-1時間點無法影響t時間點所預測的結果，若資料型態是序列數據且數據之間會互相影響的情況下，將無法合理預測當前結果，僅根據當前輸入神經元經過計算後輸出，而輸出結果中和之間並沒有連結關係，循環神經網路欲得出時間點t的輸出神經元，需根據上一個時間點t-1的輸出神經元與當前輸入神經元計算而得出，針對t時間點已存在的狀態，以表示在t-1時間點前的序列數據狀態，則表示包含序列數據狀態。

|  |  |
| --- | --- |
| 循環神經網路可分為正向、反向傳導，其正向傳導可分為兩個部分，為輸入神經元，可表示為共用權重參數，其計算公式可由參數並採用激活函數得出。 | |
|  | ( 1 ) |
|  | ( 2 ) |
| 輸出神經元經過與線上權重計算後得出，經過激活函數後得出，也可以表示為正規化後的預測機率。 | |
|  | ( 3 ) |
|  | ( 4 ) |

輸出層至隱藏層計算過程，以公式( 5 )、( 6 )表示，隱藏層至輸出層過程以公式( 7 )、( 8 )表示，循環神經網路針對數據序列執行順序分為輸入層、隱藏層、輸出層描述，其運作原理與傳統神經網路類似，可建置多層式模型架構，最大的不同點在於循環神經網路在反覆訓練中不將當前訓練參數釋放丟棄，將其參數視為共用參數，多次循環接採用相同參數，以達到記憶特性，讓序列數據能夠受到時間、排序影響，在當前時間點輸入層含有上一個時間點的輸出層神經元，其中所採用的是共用參數，每執行一次產生誤差後，根據每一個權重分配神經元相同比例的誤差值，並在倒傳遞時採用梯度下降更新共用參數，以圖 2.1循環神經網路概念圖表示。

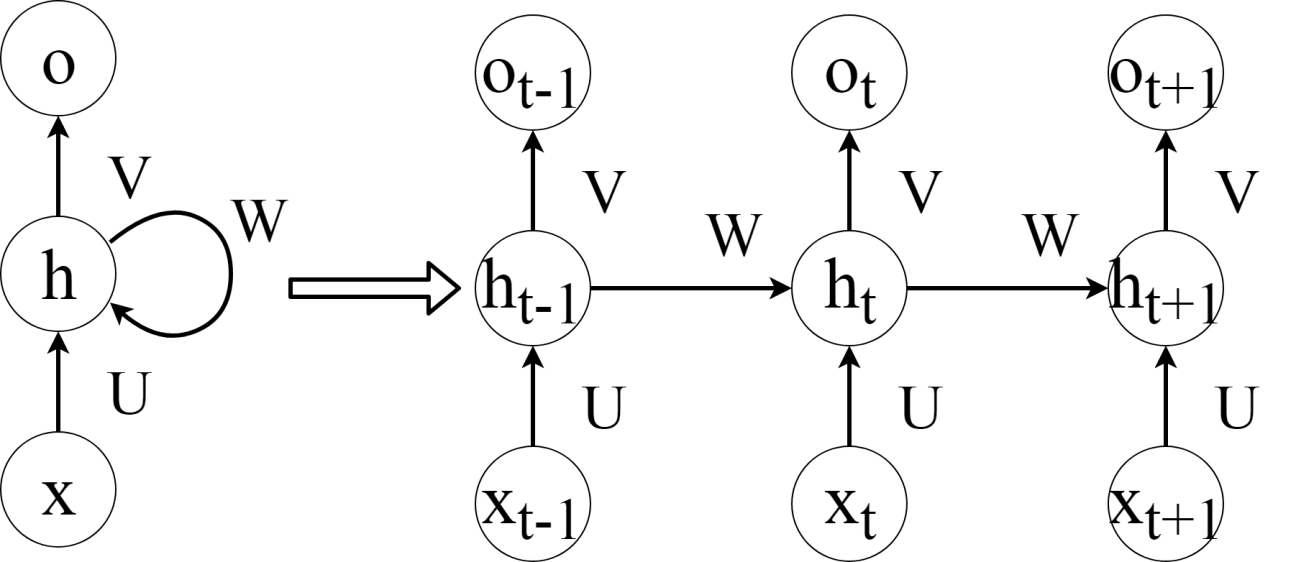


圖 2.1循環神經網路概念圖

循環神經網路採用倒傳遞演算法的方式進行誤差值修正，以梯度下降法計算權重比例調整誤差分配，由於每更新一個權重將乘上一次共用參數，若共用參數大於1經過更新後，則會產生無窮大的誤差，若小於1則導致誤差梯度隨著時間以指數般消失，也就是梯度消失，到某一個時間點後，將無法再紀錄過往的數據特徵，為了解決這個問題，Hochreiter & Schmidhuber(1997)提出LSTM，來改善原先RNN上述的問題。

### 長短期記憶(Long Short-Term Memory, LSTM)

LSTM與RNN架構相似，其中最大的不同點在於長短期記憶多出三個門管控每一個數據是否能夠通過門到下一階段，這三個門分別為輸入門、遺忘門、輸出門，數據透過sigmoid計算後，可以得到多少數據能夠通過門，可以限制數據總量，其機率範圍介於0至1之間，0表示沒有任何數據能夠通過，1表示所有數據皆可通過，在更新細胞狀態與數據狀態流程由圖 2.2和公式(5)至公式(10)依序介紹。

對於LSTM的應用與描述，其中Huang (2017) 為了提高藥物相互作用提取藥物相互作用（Drug-drug interactions, DDIE）的性能，提出了一種新方式，基於LSTM的分類器將正面實例分類為特定類別，其實驗結果表明，此模型採取的兩階段流程較以往的一階層方式效果佳，針對短期記憶與長期記憶，Cao(2018) 短期記憶被用來儲存來自最新環境中適度的解決方案，而長期記憶是儲存以前所有環境中發現的最佳解決方案，檢測到環境變化後，群體中的某些粒子位置被短期記憶成員所取代，並且在當前環境下保留長期記憶中最好的解決方案。

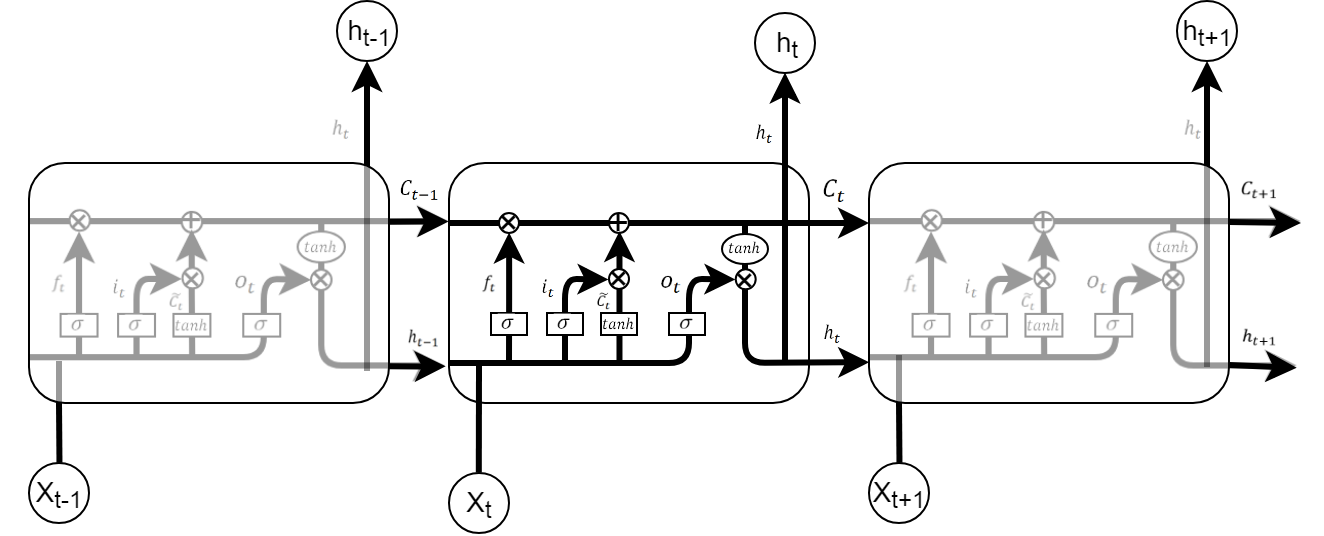


圖 2.2 LSTM結構圖

表示輸入數據經過門後所能留下多少數據，依照RNN架構中，輸入數據包含t-1時間點的狀態數據和t時間點的輸入數據，經過遺忘門計算要有多少數據保留，此公式由公式( 5 )表示。

|  |  |
| --- | --- |
|  | ( 5 ) |

公式( 6 )中表示是用於計算t時間點數據的通過概率，也是一個門的通過概率，而公式( 7 )所計算的則表示t時間點所擁有的數據總量，其數據透過激活函數計算後所得。

|  |  |
| --- | --- |
|  | ( 6 ) |
|  | ( 7 ) |

t時間點的細胞狀態以表示，其計算經過通過概率與t-1時間點的細胞狀態相乘所得到t-1時間點的數據通過量，t時間點數據總量與通過概率相乘，得到t時間點數據允許更新狀態的數據，公式( 8 )表示。

|  |  |
| --- | --- |
|  | ( 8 ) |

而在輸出門中所計算概率由公式( 9 )表示，是一個通過機率與相同概念，其不同點在於共用參數不同，上述公式計算而得到的t時間點更新狀態後的數據經過tanh激活函數和概率所得到，可以表示為更新t時間點數據後的狀態，其計算以公式( 10 )表示。

|  |  |
| --- | --- |
|  | ( 9 ) |
|  | ( 10 ) |

### 斷詞

Fxsjy(2017)開發Jieba中具有幾種參數可以變更模型執行的模式，是否啟用隱馬可夫鏈(Hidden Markov Model, HMM)、是否啟用全模式。預設模式為不使用HMM且不使用全模式，以最基礎的方式斷字，其中使用全模式時，將句子中所有可以成為詞語都掃描出來，而不使用則會將當前句子精確的斷字，此模式對於句子來說沒有多餘的字詞，適合直接文本分析，可選擇最大字數斷詞，1-Gram表示最大字數為1個字、2-Gram表示最大字數為2個字，以此類推，過去文獻也採用N-Gram等說法表示無上限字數，則是根據過去數據統計的頻率計算出現機率和權重動態選擇最優的斷字字數。

在Jieba斷字中會將每個字以四個代號表示，標記出現的機率是採用過去數據出現的頻率統計後計算而得出，訓練期間經過原始數據統計，並建立歷史數據庫作為往後斷詞步驟的依據，每筆資料都將更新數據庫重新計算頻率和權重，每次訓練數據經過計算後得出結果，也將其原始數據統計和更新歷史數據庫，測試數據可選擇是否使用隱馬可夫鏈分析標記，若識別數據為未知詞則將其標籤初始化為S，Jieba標籤代號可分為B代表詞語中的起始字、M代表詞語中的中間字、E代表詞語中的結尾字、S單字成詞，在標記代號時，ㄧ段句子經過初始化後，第一個字輸出標籤為B、S的機率很大，M、E輸出的機率近乎0，標籤根據過去文本與建立的數據庫所影響，第二個字在標記時會考量前一個字的標籤，若前一字標籤為S時，則M、E輸出的機率近乎0，相反的B、S的輸出機率較大，經過大量文本訓練後，數據庫涵蓋量變大，未知詞越來越少，能夠達到斷詞的效果。

### 詞向量(Word2Vec)

在傳統文字探勘中須將文章、文本、一段句子斷字，擷取出單詞文字後的高階語言在轉換為機器語言、向量、二位元編碼等等數值編碼方式，編碼後的字詞才能進行機器學習、深度學習、語意分析等電腦計算相關的研究，Musoll(1998)將One-Hot Encoding方式應用於工作區的編碼，能夠有效減少處理的程序。由於詞向量在傳統編碼方式過於繁瑣，耗費太多時間編碼，為了減少處理程序，本研究採用One-Hot Encoding方式，如圖 2.3所示。

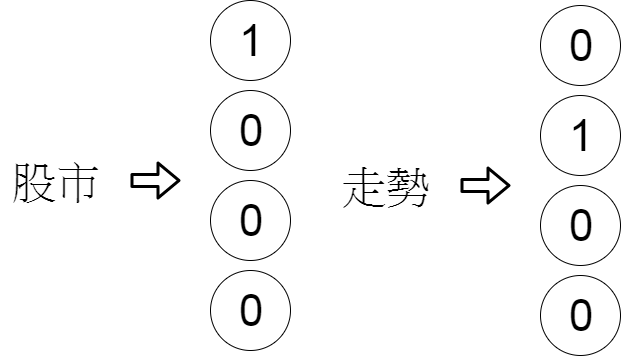


圖 2.3 字詞執行數值編碼

在One-Hot Encoding方法中能夠解決字詞無法進行運算的問題，由於編碼採用二位元表示單詞編碼，會有稀疏矩陣的問題，其作法必須消耗大量的計算資源，且需要較長的處理時間，為了避免字詞過多導致重複編碼，起初設計時需考慮文本字詞相異程度，設置可編碼字詞上限數，就現行階段越來越多非結構化文本，過往設置的上限數往往會不足以表示現階段文本字詞數，因此必須要不斷的加大字詞上限數或擴充可表示的位元數，更新後必須將以往表示的字詞一併更新，才不會與新編碼數據衝突，耗損大量時間成本將無可避免。

為了解決上述One-Hot Encoding不斷更新上限數和更新、分析耗費大量時間成本等問題，新興方式將採用Word2Vec的方式，以研究人員自行定義文本所需總字詞數量，並計算需要多少維度向量表示單一字詞，Word2Vec的做法是以類神經網路經過輸入數據計算後，擷取其中壓縮過後的語意向量，如圖 2.4所示。

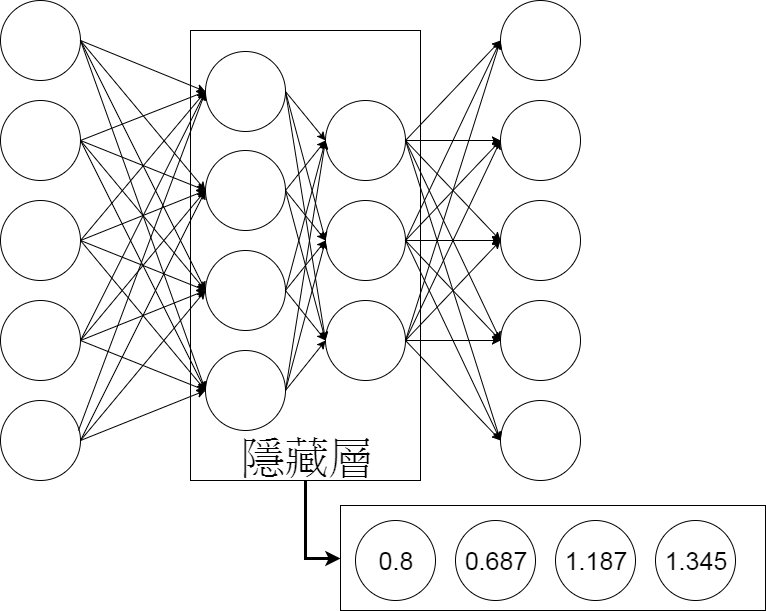


圖 2.4 Word2vec類神經網路擷取特徵向量，單詞以四個維度表示為例

Word2Vec與傳統的類神經架構相同，擁有輸入、隱藏、輸出層，在文本分析的Word2Vec模型中，句子含有的字詞經過數據編碼，再將類神經模型權重初始化，即可放入數據編碼後的字詞進入類神經網路中壓縮，若S==()表示由N個字詞按照順序構成的句子，依照統計語言模型計算某特定句子的概率可表示為P=的聯合機率，V表示為N個字詞在S中所出現的機率和前後字詞出現機率計算所得出的值，V則為各個字詞壓縮後的特徵向量，S由字詞所構成，這些字詞同時出現的機率，也將P視為壓縮語意向量，P可表示為S的句子特徵，在二維平面空間中，即可使用Cos計算兩詞的相似度，若兩詞之間的相似度很高，表示兩詞有很強烈的相關性，進而可以製作相關的應用，由圖 2.5 表示訓練時，根據字詞在句子中出現的順序排列神經元位置，輸出原數據且相同排序，其中隱藏層能夠透過結果和權重計算還原回輸入數據，在輸出和輸入層數據相同的情況下，唯一變動的僅有隱藏層參數，因此將隱藏層參數視為該句子的特徵。

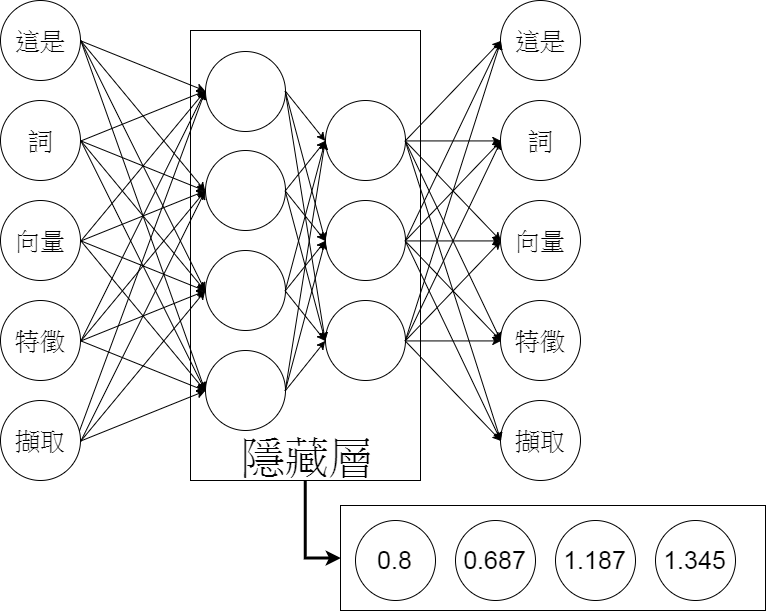


圖 2.5 Word2Vec類神經網路句子模型

Mikolov et al.(2013)表示Word2Vec的模型可分為Continuous Bag of Words (CBOW)和Skip-gram兩種架構。w(t)可表示為t時間點的字詞、w(t-1)則為前一個字詞、w(t+1)為下一個字詞，而CBOW架構是以w(t-2)、w(t-1)、w(t+1)、w(t+2)為輸入，預測w(t)為目標，Skip-gram架構則是相反，以w(t)為輸入，預測前後兩個詞為目標，以圖 2.6表示兩種模型架構。

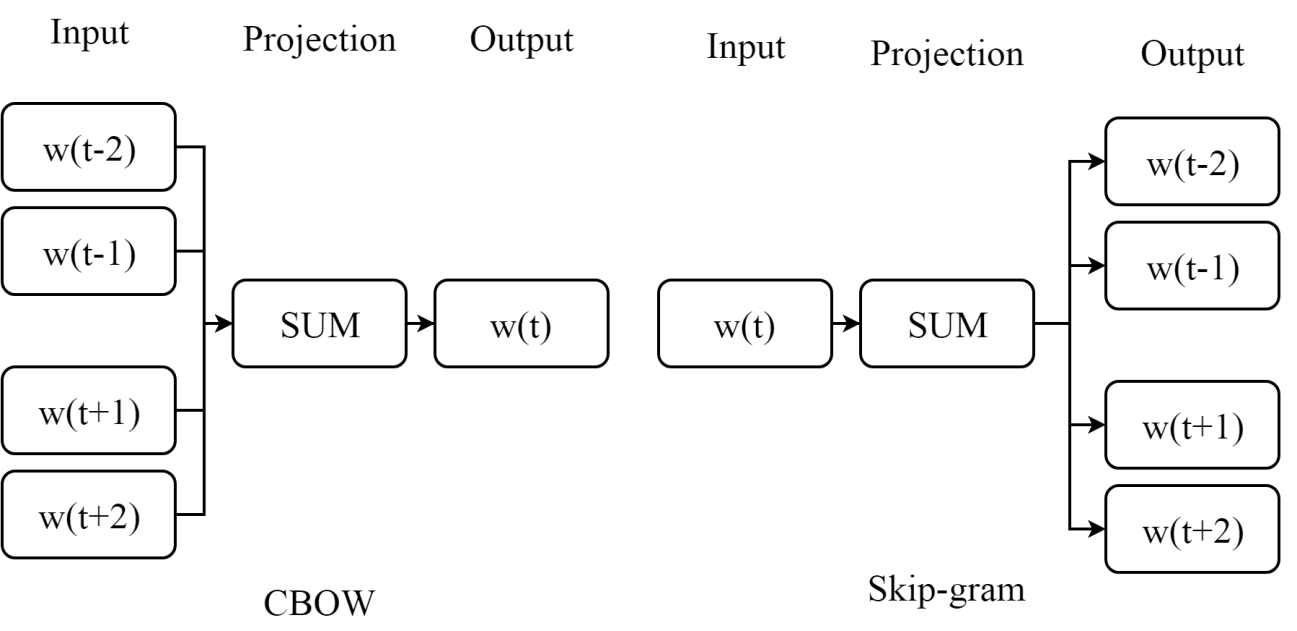


圖 2.6 詞向量訓練架構

### 關鍵詞判定

#### 2.7.1傑卡德相似係數

Jaccard(1902) 提出傑卡德相似係數(Jaccard similarity coefficient) 在兩個集合中交集範圍佔聯集的比例並計算相似係數，以統計中集合的概念計算相似程度。在文檔中兩個詞彙在所有文檔中所佔的比例為兩個集合，利用上述概念計算兩集合的相似係數，兩個詞彙其中一詞為主題詞彙，另一詞為文章中出現的詞彙，將兩者相似度做排序，提取較相似的詞彙作為該文章關鍵詞，傑卡德相似係數用來度量兩個集合之間的相似性，它被定義為兩個集合交集的元素個數除以並集的元素個數，以公式( 11 )係數計算。

|  |  |
| --- | --- |
|  | ( 11 ) |
|  | |

#### 2.7.2 PMI

Church(1990) 提出PMI(Pointwise Mutual Information)演算法，主要透過文檔中一起出現的頻率或詞彙與詞彙之間共同發生的頻率計算相似程度。使用於權重計算中，可以將文章中最具代表性的詞彙作為主題詞彙，每個詞彙與主題詞彙比較相似度，以公式( 12 )表示兩個詞彙相似度計算方式，依據此數值排序，數值越大者與主題詞彙相似度越高，也表示該詞彙對於主題詞彙越高的相關性。

|  |  |
| --- | --- |
|  | ( 12 ) |
| 語意導向PMI (Semantic Orientation-Pointwise Mutual Information, SO-PMI)，可將公式用於語意分析中的極性排序，將主題詞彙分別更換成文章中極性程度最極端者後，正向極端詞彙稱做Pword、負向極端詞彙稱為Nword，文章中每一個詞彙與Pword與Nword計算詞彙相似度，可得到兩個數值，若任務為判定詞彙極性，則將兩個數值計算，以公式( 13 )計算文章中某一詞彙*(w)*的數值，使用公式( 14 )表示作為判定依據。 | |
|  | ( 13 ) |
|  | ( 14 ) |

|  |  |
| --- | --- |
| 若以文章極性為任務，則將文章中所有詞彙的SO-PMI數值加總，若該數值大於0則表示此文章偏向正向性質，數值等於0表示文章沒有特別偏向，數值小於0此文章偏向負向性質，以公式( 15 )表示以PMI方式判定文章極性。 | |
|  | ( 15 ) |

#### 2.7.3 TF-IDF

TF-IDF是當前常見的統計計算方式，方法概要可分為兩個部分，依序分為TF(Term Frequency)、IDF(Inverse Document Frequency)，其中TF已由Jones(1972)探討並提出相關概念以及證實詞頻重要性且用數據集表示其結果，而IDF由Robertson(2004) 提到其中的計算方式，後續學者將兩者結合應用，像是Wu(2008) 建立關於訊息探索的統一標準、開發TF-IDF相關術語權重。後續學者計算特定詞在特定文章中出現的頻率次數與特定詞在所有文檔所占的比例做計算，能夠得到較以往的統計量更客觀的結果，考慮到當前文本的權重和不偏重某一篇文本。

|  |  |
| --- | --- |
|  | ( 16 ) |
| 代表文章中出現的第i個詞彙，文檔中第j篇文章，是在的出現次數，而則是在中所有字詞的出現次數之和。 | |

|  |  |
| --- | --- |
|  | ( 17 ) |
| D是所有文檔的總數，分母代表在中出現的文檔數，若為未知詞，將會導致分母為零，因此一般情況下使用+1。 | |

|  |  |
| --- | --- |
|  | ( 18 ) |
| 根據公式(1)、(2)計算出即是之權重。 | |

上述方法中，本研究探討原理後，TF-IDF最適合使用於各詞彙計算重要程度，傑卡德相似係數根據集合計算兩詞彙相似度，研究中並沒有定義每一篇文章的主題詞彙，也無法判定所有文章所對應的標籤或，在PMI演算法中，根據詞頻計算相似程度，仍然會有相同的問題，兩個詞彙互相比較相似度，且與傑卡德相似係數相同無法先行定義每一篇文章的主題詞彙，在這兩種算法中較不符合本研究分析的結構，本研究選擇TF-IDF的特徵擷取方式是因為在IDF算法中，將特定詞佔所有文檔個數比例考量計算，較不會因為單詞在某一篇大量出現而影響權重，且較符合本研究資料結構，TF-IDF特徵擷取方式本研究分為兩種，第一種屬於TF-IDF原有的概念一個文檔為一個粒度，第二種為一天為一個粒度，採用相同概念，不同粒度的方式做計算。

### 股市應用相關研究

喻欣凱(2008)使用SVM(Support vector machine)、Bayesian、C4.5預測分析股價漲跌，並使用混亂矩陣評估優劣。其中資料使用聯合知識庫內的經濟日報與個股相關之新聞資料，其股票選擇宏達電、台積電、正新三檔股，收集期間於2007年1月1日至2007年12月31日止，其數據特徵包含個股開盤變化量、大盤指數開盤變化量等等技術指標做為數據，共有19個，將個股新聞文章經過中研院斷字系統(Chinese Knowledge and Information Processing, CKIP)、TF-IDF擷取關鍵字詞等等預處理。對於文章特徵，設定單一門檻值擷取出來的特徵詞彙作為文章的標籤，並移除TF-IDF過低的詞彙作為文本特徵，目標欄位由每日收盤變化量是否有超過門檻值k為標準，可分為上漲、下跌、持平三個類別，排除一些研究限制後，使用機器學習的方式分析評估，並比較各分析方式優劣。

李春淋(2010)透過文章新聞與數據預測股價漲跌，並評估對於各類股投資報酬率。資料由2009年4月8日至2009年9月29日，共有600筆(週)股票漲跌、新聞資料，其中分為電子、金融、營建、鋼鐵四類股，共有24種股票，將各個股票的每日相關新聞，按照該新聞可能對個股股價產生之上漲下跌影響，將每日的新聞量化成「很好」、「好」、「普通」、「差」、「很差」五個等級，計分依次為2分、1分、0分、-1分、-2分。而「很好」是指可能新聞對股價會引起股價大漲；「普通」是指新聞對股價沒甚麼影響；「很差」是指該新聞可能會引起股價重挫，在數據上分為新聞、類股、大盤、市值，並對資料進行敘述性統計、交叉表、卡方檢定、變異數分析、多重比較、相關與回歸等統計分析。此篇研究中，個股新聞除了與類股、市值、大盤漲跌有顯著相關外，「好」新聞比「壞」新聞之個股股價平均漲幅增加2.80%。所以單考慮新聞而言，如果各大報紙對於單一個股所報導的好消息越多，投資者買進該股，經過一段時間後，該股票股價上漲的機率會比較高，投資大眾獲利的機率也會比較大。

江旻緯(2014)使用股價資料經過統計計算後，預測股價未來走勢。將台積電、正新、宏達電、鴻海四類股，由TEJ蒐集股價資料，由2003初到2012年底共十年期間，其中2003年至2009年為訓練資料，2010至2012為測試資料，將數據經過乖離率 BIAS、相對強弱指標(Relative Strength Index, RSI)、威廉指標(Willams’Oscillator)、心理線(Psychology line, PSY)、隨機指標（Stochastic Oscillator, KD）轉換成各類別數據做為特徵，投入五個機器學習演算法中訓練預測，並比較其結果。研究中實驗個股的延遲天數、上漲下跌之幅度的門檻值，並評估訓練預測四類股票精準度。

謝政翰(2017)利用歷史股價和四種演算法預測分析，混亂矩陣用於評估精準度。歷史資料來自TEJ蒐集，期間由2014年10月13日至2016年12月29日共546筆，包含80%訓練資料與20%的測試資料，其中開盤價差、最高價差、最低價差、RSI、股價漲跌等等共19個指標作為數據特徵，各特徵指標採用機器學習中四種演算法分析預測結果，其結果決定各個股票中最重要的數據指標，將排名前5的數據特徵經過隸屬函數切區段分析，模糊隸屬函數分六種方式將所有數據規範在同一尺度下，作為演算法的輸入特徵，經過分析預測出股市的漲跌。

由上述股市應用相關研究文獻中可以發現，進行股價分析預測，多半採用一般股市數據經過許多金融分析曲線製成不同的數據特徵，最後使用混亂矩陣做為評估指標，可以說這些文獻都在探討兩個要點，當前階段金融分析曲線對於股價是否擁有預測的能力，技術分析線達到某一個門檻值對於未來是否產生既定結果。其中李春淋(2010)的實驗模式與本研究相似，差異在於新聞文章處理方式與獲取數據的不同，且在本研究中對於文章計算極性分數，探討極性詞庫和文本分析的重要性，並結合過往文獻所使用的數據特徵，加上LSTM保留過往特徵的特性，加強數據對於分析結果的影響程度。

以收集時間的觀點中，上述文獻資料收集期間在2年內居多，且1年的資料會讓模型更完整擬合，每一年資料特徵都有所不同，完全擬合1年的資料太過主觀，當進行2年以上的分析預測時，擁有2年的資料也可以表示有兩次的循環，以年份區隔可將特徵分為兩部分，對於資料涵蓋年份越多，模型所具備的特徵相對客觀，較不容易受到其中個案影響，因此本研究採用2年的方式分析股價、指數較為客觀。

## 實驗流程

# 

財務股市相關資料中可獲得的資料不外乎是新聞文章、每日數據變化、技術分析數據，一般投資者會觀看分析資料推估未來是否看好，使用上述資料證實未來受到這些因素影響，研究流程可由資料收集、文本特徵、數據特徵、長短期記憶分別做介紹，其整體實驗架構由圖 3.1表示。

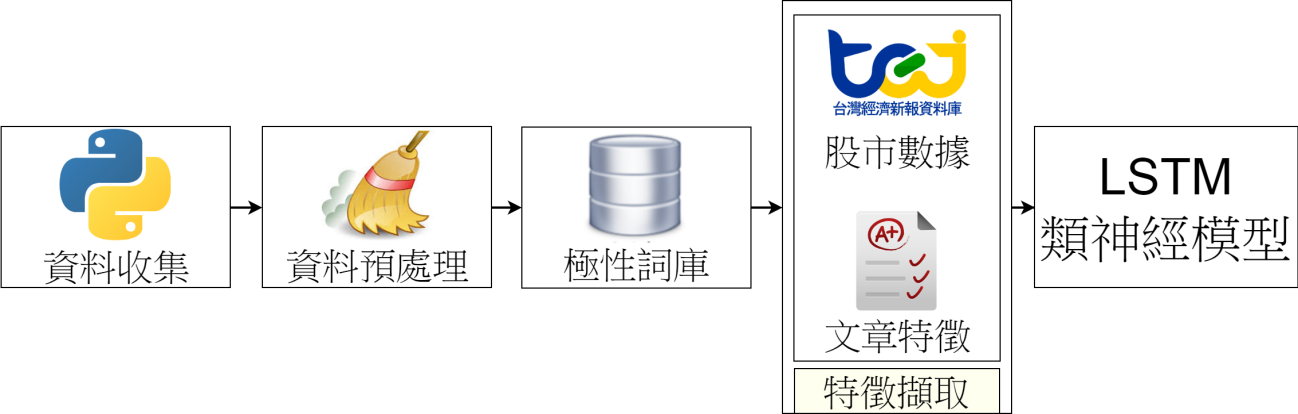


圖 3.1 研究流程圖

### 資料描述

從Google收集個股相關的新聞文章，本研究以台灣50相關新聞為例，來源有中時理財、Yahoo股市新聞、Google財經新聞、中國電子報台灣常見的新聞網，從2016/01/04至2017/12/29期間內相關的文章皆為實驗資料範圍，台灣50相關新聞文章共有每天蒐集25篇文章，共有95萬筆文章，由於歷史資料有缺失加上各個電子新聞報導相同事蹟，經過與處理後剩下13萬筆文章，後續研究流程計算文章特徵作為輸入數據，目標分為漲跌兩個類別，依據日期和Lag由文章中符合日期區間者做為預測日期的訓練資料，訓練與測試切割比率為8:2，訓練筆數為391筆、測試筆數為98筆，則訓練時間由2016/01/04至2017/08/10，測試時間為2017/08/11至2017/12/29，個股相關數據由TEJ收集，一筆資料中包含多筆文章判定極性後的數量，在金融證券市場中，TEJ能夠提供基本分析所需要的資訊，到目前為止已經成為台灣知名財經數據提供者，對於研究者來說，許多資料可以由此獲得，且具有規模與公信力，數據收集期間也與新聞收集期間相同。本研究訓練時會設定延遲時間，Lag-1表示遞延天數為一天，則使用t-1時間點的所有文章預測t時間點時，Lag-2表示遞延天數為兩天，則使用t-1至t-2時間點的所有文章預測t時間點時，獲得區間內的所有文章後，經過後續步驟所描述的細節依序處理。

### 實驗流程

#### Step1-資料收集

採用python的方式收集新聞資料，在網路文章中含有大量的文字，此步驟進行第一次篩選，能夠有效降低數據量，因此本研究收集個股相關新聞，則以個股名稱作為關鍵字探索文章中是否含有關鍵字作為判定標準，若含有則擷取該篇文章段落。

#### Step2-資料預處理

##### Step2.1-資料清理

資料內有許多不重要的資訊，像是Javascript、HTML樣式和標籤等等，對於文本分析來說，這些都是屬於不相關的資料，若擷取過多會造成資料偏移或雜訊過多等情況，在此步驟須將不重要的資訊去除，並擷取文章中含有關鍵字的段落或句子，減少不需要的資料，提升訓練效益。

##### Step2.2-斷字

比較CKIP系統後，就擴充性、斷字速度、執行流程來說，jieba引用能夠讓整體計算與架構更為明確和順暢，因此在斷字方面採用jieba和自定義的關鍵字字典，在後續步驟能夠更簡單地找到關鍵字並擷取計算，輸出將以空格隔開字詞與字詞之間以空格區隔，輸出可分為完整文章與擷取特定詞性的關鍵字詞，本研究採用詞性輸出的方式，其資料集中含有動詞、形容詞、副詞、人名、公司名稱等等詞性，此步驟將完整輸出加上詞性，假設動詞、形容詞、副詞最有可能代表極性，而擷取名詞主要是區分極性所表示的對象，原句為：「我每節下課都開心地跑向溜滑梯」，其中擷取的是「我」、「開心」、「跑」、「溜滑梯」，擷取這些具有代表性的詞性，即可說明這一段句子所代表的極性。

##### Step2.3-過濾詞性(Filter Parts of speech)

中央研究院資訊科學所詞庫小組提出研究中(Ma & Chen., 2003;Ma et al., 2012) 表示CKIP線上系統的詞類、文法實作概念，探討中研院平衡語料庫詞類標記集詞類細項可分為46類，Jieba可分為99種詞類。上述兩種詞性標記方式皆採用階層式分類，兩種標記的方式與傳統中文、英文字典比較後可以發現，第一階層的詞類與傳統的字典較為相似，第一階層的詞類概括較多細項詞類，在此步驟進行詞性的過濾，以大項目作為過濾標準較有效率，由表 3.1表示四種來源比較不同的詞類標籤，其中七個類別是四種來源都有的詞類。

表 3.1 詞類標籤比較表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 比較項目 | 中研院  平衡語料庫 | Jieba | 中文字典  (現代漢語) | 英語 |
| 第一層  詞類個數 | 8 | 22 | 8 | 10 |
| 相同詞類 | 名詞、動詞、形容詞、副詞、介系詞、連接詞、感嘆詞 | 名詞、動詞、形容詞、副詞、介系詞、連接詞、感嘆詞 | 名詞、動詞、形容詞、副詞、介系詞、連接詞、感嘆詞 | 名詞、動詞、形容詞、副詞、介系詞、連接詞、感嘆詞 |
| 相異詞類 | 語助詞 | 語助詞、代名詞、數詞、時間詞、處所詞、方位詞、區別詞、狀態詞、量詞、語氣詞、擬聲詞、前綴、後綴、字符串、標點符號 | 代名詞 | 代名詞、冠詞、數詞 |

表 3.2的相同詞類中，分別為名詞、動詞、形容詞、副詞、介系詞、連接詞、感嘆詞共七種，在七種詞類中依據各類定義排除非具有極性詞，維基彙整不同專家、業餘等貢獻者提供具有一定的參考價值，本研究依據維基表示的定義做為參考，並判定該詞性是否具有極性，由表 3.2描述定義。

表 3.2 詞性類別定義與細項分類

|  |  |
| --- | --- |
| 詞類 | 定義 |
| 名詞 | 名詞在不同的語言中可以分為「專有名詞」、「普通名詞」和「集合名詞」等等。另一種分類方法是把名詞分為「具體名詞」和「抽象名詞」。前者指有實體的確定事物（如貓、床、金字塔等）；後者指情感、意見、概念等抽象的事物（如自由、正義等）。 |
| 動詞 | 動詞可分為一般動詞和特殊動詞，一般動詞可分為行為動詞、心理活動動詞、使令動詞、比似動詞等，特殊動詞可分為助動詞和趨向動詞，趨向動詞表示動作的趨向，可以作謂語，也可當作補語，助動詞是表示意願或可能的詞，可以作謂語。 |
| 形容詞 | 周一民 (2006)表示形容詞可分為三類，依序為：  1.性質形容詞。能受「很」修飾，表示的性質沒有一定的界限，認知有主觀性。如：大、少、美麗、可憐。  2.狀態形容詞。不能受「很」修飾，因為本身已經含有表示程度的語素。如：雪白、滾燙、紅彤彤、黑不溜秋。  3.非謂形容詞。不能直接充當謂語。如：男、副、初級、多功能。 |
| 副詞 | 用以修飾動詞或加強描繪片語或整個句子的詞，修飾名詞的詞一般為形容詞，又稱限制詞。 |
| 介係詞 | 虛詞的一種，本身沒有實在的意義，主要用在名詞、代詞前面，組成介詞結構。 |
| 連接詞 | 用來連接詞語、短語、句子、段落等的詞，表示被連接的語言單位之間的關係。連不同的邏緝關係。連接詞和詞組的連詞，跟連接句子的連詞並不相同。 |
| 感嘆詞 | 用於表達各種感情的詞。它與後面句子的其餘成分無語法聯繫。 |

由表 3.2清楚了解詞類意義，感嘆詞雖然可以表達發言者情感，但與句中無語法聯繫，表示發語者不管感嘆詞是怎樣的詞，後續的句子都能夠通順並無語法問題，在句子、文章的極性判定中排除模玲兩可的詞，在七種詞類中本研究保留名詞、動詞、形容詞、副詞四種詞類。

##### Step2.4-詞向量

將歷史資料斷字後，Python-Word2Vec套件計算每一個字詞與字詞之間的關係、相似度、權重，並以隱藏神經元的數值作為單一詞彙的多維向量，以此向量表示該詞的特徵，本研究採用預設參數100個維度表示一個詞，由於無法呈現100維度向量圖，圖 3.2以100維度中其中一個二維平面角度呈現。

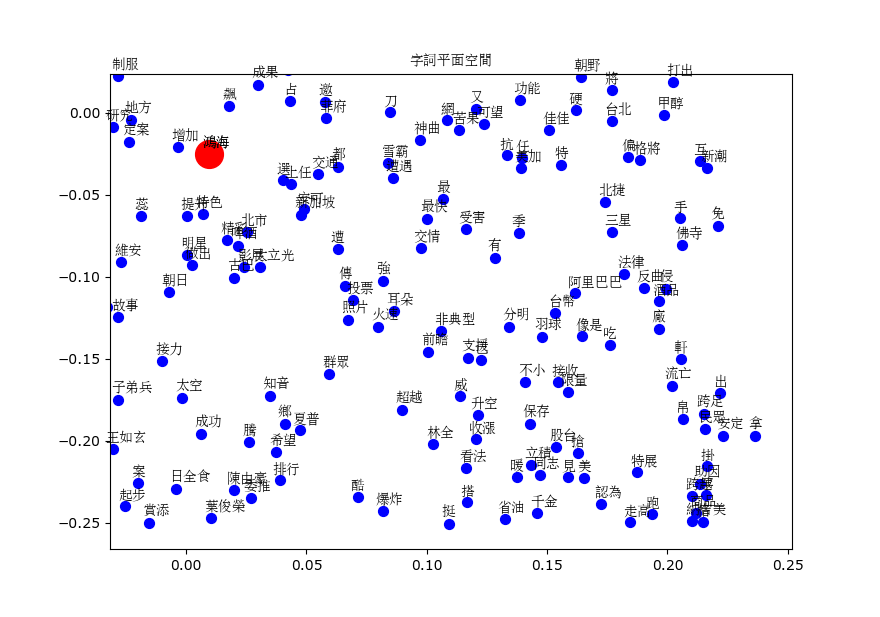


圖 3.2二維空間平面表示斷字後的向量特徵以2016/01/18，以Lag-1為例

##### Step2.5-相關詞

每一個詞用100個維度表示，再多維空間中用Cos計算每個字詞與關鍵字詞的相似性，由公式( 19 )表示，其中a屬於A詞彙向量矩陣，b屬於B詞彙向量矩陣，將A、B向量矩陣畫至二維平面中並計算兩條線之間的夾角角度，用此角度可以判定兩詞彙相似程度，其角度等於0則判定兩詞彙近乎相同，角度等於90表示兩詞彙沒有任何相關，若角度為180可以判定為兩詞彙意義完全相反，也可表明為角度越小表示越相近。

|  |  |
| --- | --- |
|  | ( 19 ) |

由於二維平面無法使用100維度表示一個詞，為了表示空間概念，圖 3.3以2個維度表示一個詞放入二維平面空間示意，資料為2016年01月17日所有文章，擷取關鍵字周圍的100字，目標欄位為2016年01月18日漲跌數據，其中紅色標記為關鍵字詞、接著依據關鍵字相關程度高到低，由顏色與大小大到小依序為，橘色、黃色、墨綠色，藍色標記為與關鍵字較不相似的字詞，2016年01月18日股價上漲，則將黃色標記詞彙增加上漲次數1次。

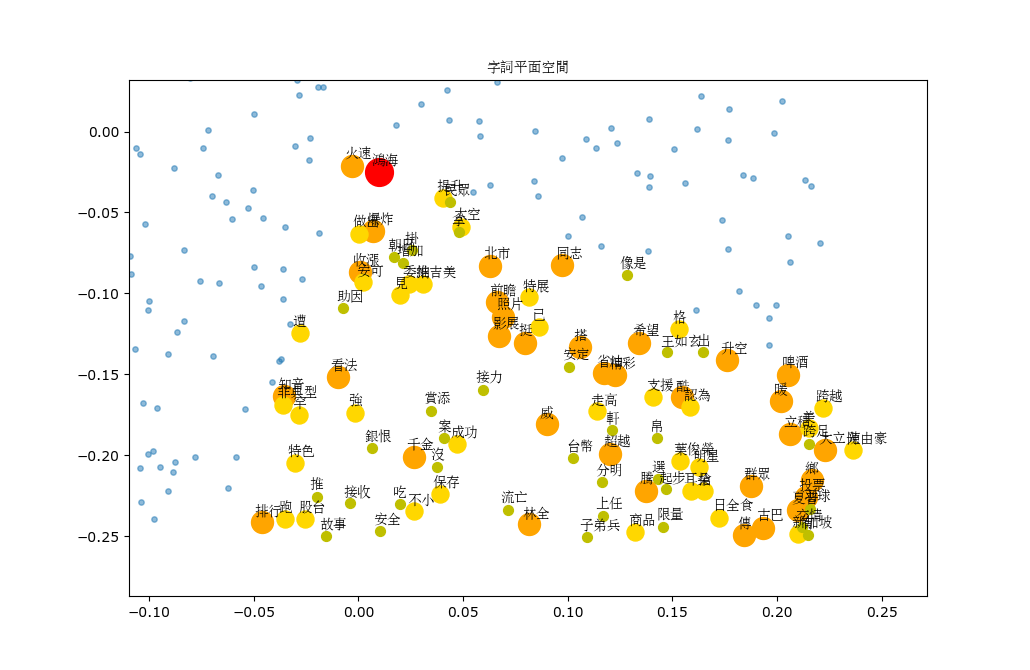


圖 3.3 以二維空間中表示關鍵字周圍的100字以鴻海新聞2016/01/18，以Lag-1為例

#### Step3-極性詞庫

從Step2.5中可以將一篇文章擷取出個股關鍵字相似排行前100個字詞，將這些字詞皆判定為漲跌的相關字，訓練目標若為上漲，則將100字視為正向字詞，最後統計單一字詞被視為上漲、下跌的天數，本研究經過訓練後擁有53356個字詞，以其中3個字詞作為表示，表 3.3 表示本研究使用部分字詞統計上漲、下跌天數數據。

表 3.3 字詞統計

|  |  |  |
| --- | --- | --- |
| 字詞 | 上漲天數(正向) | 下跌天數(負向) |
| 受惠 | 2287 | 2863 |
| 抗跌 | 601 | 681 |
| 下修 | 246 | 305 |

2016年至2017年間每日股價變化量，下跌天數較上漲天數多，將會導致極性偏斜，無法根據相關字詞統計次數判定正確極性，本研究將相關詞用TF-IDF以文檔為粒度的方式計算，並計算訓練期間的平均TF-IDF，以下依序介紹計算步驟，由公式( 20 )、( 21 ) 、( 22 ) 、( 23 ) 、( 24 )，以及最後以( Decision TF-IDF)表示的極性的公式( 25 )。

|  |  |
| --- | --- |
|  | ( 20 ) |
|  | ( 21 ) |
| *j*表示文檔數，表示第i個相關詞，(Positive TF-IDF)、(Negative TF-IDF)分別表示訓練目標為上漲、下跌的TF-IDF權重，字詞出現於目標欄位為漲的權重加總以PSTI (Positive Sum TF-IDF)作為變數名稱，字詞出現於目標欄位為跌的權重加總則以NSTI(Negative Sum TF-IDF)。、可表示訓練期間中，出現一詞的權重，經過加總以、表示，經過上述公式計算得出結果由表 3.4 呈現。 | |

表 3.4 字詞正、負向的權重總和

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 字詞 |  |  | |  |  |
| 受惠 | 2287 | 2863 | | 123.9066509 | 143.3470114 |
| 抗跌 | 601 | 681 | | 19.43024739 | 19.03809676 |
| 下修 | 246 | 305 | | 2.874792197 | 3.718900919 |
|  | | | ( 22 ) | | |
|  | | | ( 23 ) | | |
| 根據公式( 20 )、( 21 )可得、，、為一詞的訓練目標為上漲、下跌統計天數，則PATI(Positive Average TF-IDF)、NATI(Negative Average TF-IDF)分別表示在過去資料中以日為單位的正、負平均權重，本研究以、表示上漲、下跌的影響程度，對於字詞來說，出現一詞對於股價上漲的影響程度，其部分字詞由表 3.5表示。 | | | | | |

表 3.5 平均權重

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 字詞 |  |  |  |  |  |  |
| 受惠 | 2287 | 2863 | 123.9066509 | 143.3470114 | 0.05417868 | 0.05006881 |
| 抗跌 | 601 | 681 | 19.43024739 | 19.03809676 | 0.03232986 | 0.00360705 |
| 下修 | 246 | 305 | 2.874792197 | 3.718900919 | 0.01168614 | 0.01219314 |

|  |  |
| --- | --- |
|  | ( 24 ) |
|  | ( 25 ) |
| 根據公式( 24 )、( 25 )，比較與大小，若較大，表示一詞對於股價上漲的影響程度較大，若較大，則反之，根據與計算判定一詞的極性，分為正向、中性、負向三類詞性，以此訓練極性詞庫。 | |

#### Step4-特徵擷取

##### Step4.1股市數據

從TEJ獲得每日文章正、負向文章變化量做為輸入數據特徵，其變化量對於未來漲跌屬於時間相關特徵，根據不同的時間為基底產生與原有文章量不同的特徵，增加對於未來走勢的影響，其中謝政翰 (2017)，也以每日變化量做為輸入數據的想法相似，實際情況下，新聞文章若透露出個股太多的負面新聞，會間接影響股價走勢和決策者的看法過於悲觀，會過度影響該股未來走勢，由表 3.6可以看出正規化前與正規化後的變化。

表 3.6鴻海(2317) 每日數據與正規化

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 數據  日期 | 交易量 | 交易金額 | 股價  變化量 | 交易量  正規化 | 交易金額正規化 | 股價變化量正規化 |
| 2016/1/4 | 35223 | 2804319 | 2.2 | 0.177069 | 0.144211 | 0.87 |
| 2016/1/5 | 30125 | 2386189 | -0.1 | 0.146666 | 0.118673 | 0.64 |
| 2016/1/6 | 42557 | 3325714 | -0.7 | 0.220807 | 0.176055 | 0.58 |

##### Step4.2文章特徵

新聞文章在極性詞彙統計之下，可以發現正向文章普遍多於負向文章，是合理的情況，對於機器來說，資料不平衡的問題，將會過度學習較多的類別，正向文章普遍多於負向文章，則正向文章的特徵將會明顯多於負向文章，導致負向資料無法有效學習，為了擬平兩者極性文章數量的均值誤差並凸顯文章特徵，在訓練期間計算各極性的最大值與最小值，將每天文章變化量正規化，由公式( 26 )、( 27 )表示，讓神經網路倒傳遞誤差修正時能夠得到最適當的配置，其訓練期間部分結果由表 3.7所示。

|  |  |
| --- | --- |
|  | ( 26 ) |
|  | ( 27 ) |
| (Positive Decision)表示在t-1至t時間點的正向文章變化量，(Negative Decision)表示t-1至t時間點的負向文章變化量，、在訓練期間中正向文章變化量的最大、最小值，、在訓練期間中負向文章變化量的最大、最小值，經過計算後得出(Positive Decision Normalization)、(Negative Decision Normalization)可表示為、經過正規化後的數值，介於0至1之間。 | |

由Step3可以得到每一個詞與股價上漲、下跌的影響程度，每篇文章中含有許多字詞，由極性詞庫能夠獲得每個字詞的極性分數，由公式( 24 )計算出表示，根據每個相關字詞的極性分數計算文章分數，如公式( 28 )表示。

|  |  |
| --- | --- |
|  | ( 28 ) |
| AC(Article Count)為特定文章中出現字詞的次數與做計算，得出作為文章特徵，根據作為判定文章極性重要指標，為正值則文章為正向、為負值文章則為負向。在Lag期間擁有m篇文章，每個字詞能夠獲得由極性詞庫中採用TF-IDF公式計算的，在t時間點中涵蓋在時間範圍內的所有字詞，根據詞庫獲得並經過計算得出文章極性分數，由公式( 28 )計算得出，依據公式( 29 )函數判定特定文章極性。 | |
|  | ( 29 ) |

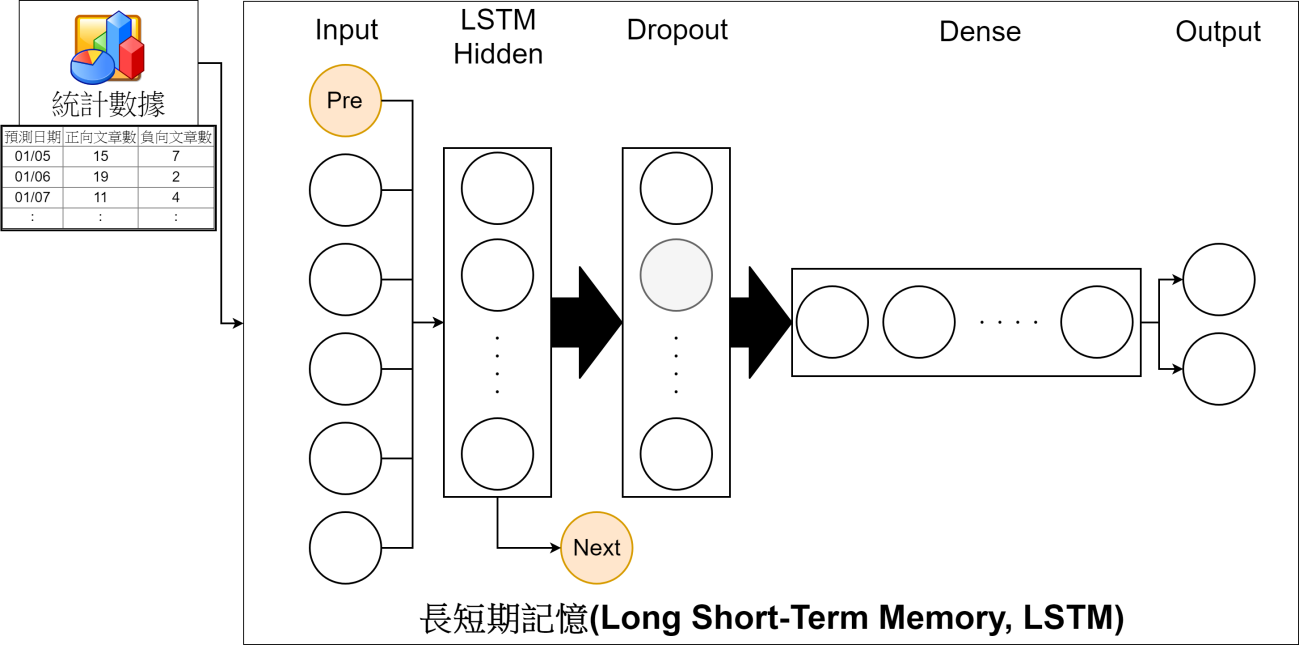
以作為判定正、負文章的門檻值，經過統計次數所得到某特定時間點出現正、負向文章數量並計算每日文章數量與變化量，由判定為0則該文章屬於中性文章，代表對於未來上漲、下跌沒有任何的影響，因此不考慮中性文章。文章數量每日的變化量正規化，以表 3.7所示。我們可以發現負向的變化量，經過正規化後會發現數值較小，輸入至模型時，數值較小相對的影響較小，將會抹煞負向特徵，因此本研究在文章數量變化量不採用正規化方式計算。

表 3.7 每日正、負文章量，以鴻海為例

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 數據  日期 | 正向  文章 | 負向  文章 | 正向文章  變化量 | 負向文章  變化量 | 正向文章變化量正規化 | 負向文章變化量正規化 |
| 2016/1/3 | 10 | 5 | 0 | 0 | 0 | 0 |
| 2016/1/4 | 15 | 7 | 5 | 2 | 0.580645 | 0.615385 |
| 2016/1/5 | 19 | 2 | 4 | -5 | 0.548387 | 0.076923 |
| 2016/1/6 | 11 | 4 | -8 | 2 | 0.16129 | 0.615385 |

#### Step5-LSTM

根據RNN所具有的特性，能夠保留先前數據中重要的特徵，其中RNN造成梯度消失與梯度爆炸的問題，而LSTM對於此問題擁有解決方式，且經過長期記憶通道、遺忘門、激活函數的計算，保留比RNN更久遠的記憶。以金融觀點來看，今日的個股股價會受到前1天、2天甚至是3天以上的數據影響，且股價與自身具有相關性，根據數據特性，採用RNN與LSTM的時序分析方式較合適，為了保留更完整的長久記憶，LSTM作為本研究分析評估方式。

****圖 3.4 LSTM架構圖

##### Step5.1-Input

在t的時間點時，輸入數據共有6個，第1個數據為*Pre*的神經元，其數值由t-1時間點所有輸入數據經過遺忘門、激活函數、權重計算後所得出的數值，其中2個神經元由、代表，可表示為正、負向文章變化量正規化後數值，剩下的3個神經元數據皆來自TEJ數據，若以個股為例，其數據則預測日期前一天的股價變化量、交易股數、交易金額，交易股數及交易金額若計算每日變化量，長久記憶後將會遺失特徵，導致門檻值判斷誤差，因此本研究並沒有採用變化量作為數據特徵，簡而言之，若數據第一天到第十天以100股的數量遞增，至第十天為1000股的交易股數，假設該股1000股的交易股數已經屬於相當熱門，但根據變化量僅能看出與前一天的差距100股，並不能表示第十天的交易股數是具有一定熱門程度。

表 3.8 LSTM輸入數據，以鴻海Lag-1為例

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 預測  日期 | 數據  日期 | 正向文章  變化量 | 負向文章  變化量 | 交易量  正規化 | 交易金額  正規化 | 股價變化量  正規化 | 漲跌 |
| 2016/1/5 | 2016/1/4 | 5 | 2 | 0.177069418 | 0.144210702 | 0.87 | 漲 |
| 2016/1/6 | 2016/1/5 | 4 | -5 | 0.146666269 | 0.118673134 | 0.64 | 跌 |
| 2016/1/7 | 2016/1/6 | -8 | 2 | 0.22080749 | 0.176055249 | 0.58 | 跌 |

##### Step5.2-LSTM hidden

輸入數據經過線上權重計算後放入LSTM隱藏層中，其架構採用原生LSTM設定，依序為循序型(Sequential Dropout)、循環型(Recurrent Dropout)，兩者主要差異點在於用的位置不同，Sequential Dropout用於輸入至隱藏層之間、Recurrent Dropout用於第一個時間點輸出的特徵至第二個時間點的輸入之間，由圖 3.5中A、B表示兩種型態的Dropout，輸出神經元根據不同的資料型態會有所不同，計算的變數也會有所不同，造成結果因此而受到影響。

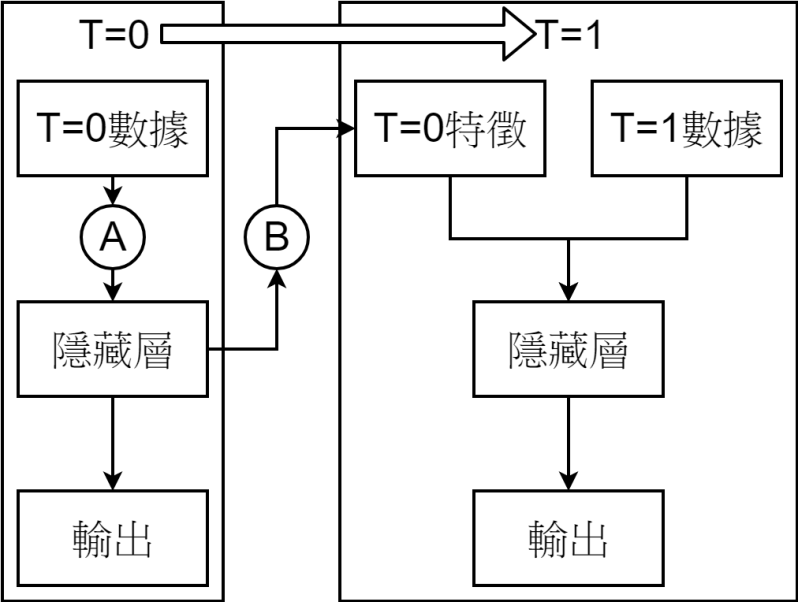


圖 3.5 LSTM Dropout位置圖

##### Step5.3-Dropout

將Step5.2-LSTM隱藏層輸出至Step5.4-Dense前的神經元數值遮罩特徵，Dropout的數值介於0到1之間，Srivastava et al. (2014) 提到Dropout對於每個隱藏層神經元、每一個迭代次數都是獨立情況，若Dropout數值為0.5的情況下，每一次迭代每一個隱藏層神經元都有50%的機率被遮蔽，數值設定過高導致神經元特徵完全被遮蔽，訓練集特徵將無法被模型學習，過低的情況下可能會導致模型過度擬和訓練集，測試集和訓練集特徵差異過大加上過度擬和訓練集，導致測試結果不理想，Dropout可以改善過擬和的情況，設置的位置與大小也會是一個很關鍵的因素之一，本研究在隱藏層與全聯階層之間安插一個Dropout層(Dropout Layer)，成為後續時間中一個參數，藉此證明位置與數值大小和測試結果是否有密切關係。

##### Step5.4-Dense

在batch或Lag大於2的情況下，2筆以上的資料作為一綑，在訓練所使用的資料編排格為2維表示，執行類神經分類任務時所需要的格式需要1維陣列的資料格式，加入此層可以將2維資料攤平為1維資料，正好接續後續的分類任務，此層表示為一般常用的全連結層，可以稱為密集層，用圖形來看連接最密集的階層就在這一層，由圖 3.6表示轉後前後的位置，轉換前為2x4矩陣，轉換後為1x8的矩陣。

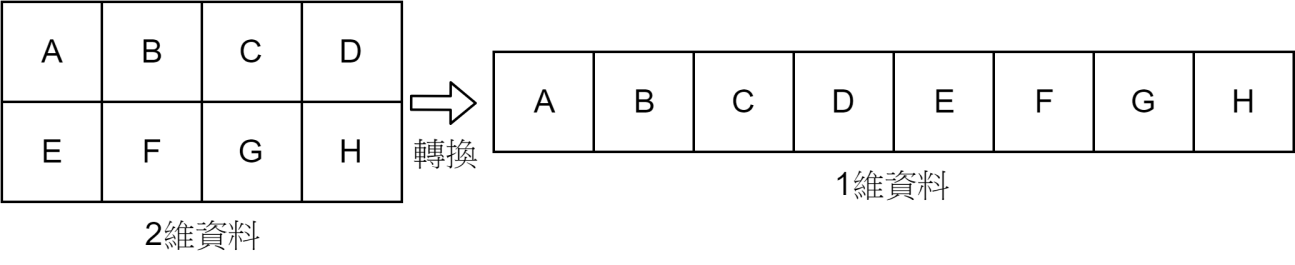


圖 3.6 2維資料攤平成1維資料

##### Step5.5-Output

結果輸出分為漲跌兩個類別，個股根據每日開盤價至收盤價的變化量為基準，台股大盤則以開盤加權指數至收盤加權指數的變化量為基準，若變化量大於等於0則視為該天上漲、若小於0則視為該天下跌，此類別採用one-hot encoding的方式表示類別，可以放大目標欄位特徵，增加學習成效。

## 實驗設計與結果

# 



### 個股(三類股)台積電、鴻海、台塑化

以市值占大盤比重排名前三名作為參考依據，台積電2016年佔18.26%、鴻海佔6.47%、台塑化佔3.23%，以此三類股作為本研究實驗個股對象，其資料由2016年初至2017年底共兩年期間。經過上述實驗流程一步一步進行處理並輸入模型分析，透過滑動視窗(Sliding Window)的方式進行序列數據擷取，依序丟入模型中，透過試誤法(Trial and Error)調整至最佳參數後，進行延遲天數(Lag)、迭代(Epoch)與每一檔股票的關係，以台灣大學公開的極性詞庫(NTUSD)與本研究的各別實驗，後續章節依序介紹個股與大盤的研究設計與結果呈現。



#### 本節實驗目的

* 採用本研究流程對三種不同市值的個股實驗，是否能夠達到一定的水準？
* 新聞反應的時間對於三種不同市值的個股是否具有影響力？
* 新聞反應至個股最佳天數為何？

#### 台積電

使用本研究詞庫分析台積電新聞的LSTM架構中，隱藏層神經個數(LSTM Units)使用256個神經元、激活函數(Activation)使用sigmoid、Sequential Dropout設置為0.1，也就是在輸入至隱藏層的過程中，每一個神經元都有10%的機率被模型遮蔽、Recurrent Dropout其放置的位置在第一個時間點至的二個時間點之間，與Dropout的功用相同，其數值設置為0、誤差函數(Loss Function)採用均方差(Mean Squared Error, MSE)評估實際值與預測值的誤差並修正，學習率(Learning Rate)設置為0.001能夠有效的學習台積電新聞文章數量特徵，Dropout層 (Dropout Layer)是放置於LSTM隱藏層至Dense層之間，在台積電新聞文章分析設置為0效果最佳，與其他優化器比較後，決定使用adam演算法優化較為適合。上述設定由表 4.1表示，其參數皆為本研究經過Trial and error後獲得各個參數最佳值，Lag的表現中，由表 4.2可看出延遲3天所預測的精準度最高，由此可知台積電在新聞分析中，Lag設為3天最為適當。

表 4.1使用本研究詞庫分析台積電新聞所使用的參數

|  |  |
| --- | --- |
| LSTM Units | 256 |
| Activation | sigmoid |
| Sequential Dropout | 0.1 |
| Recurrent Dropout | 0 |
| Loss Function | MSE |
| Learning Rate | 0.001 |
| Dropout Layer | 0 |
| Optimizers | adam |
| Epoch | 100 |

表 4.2使用本研究詞庫和不同Lag分析台積電新聞

|  |  |  |
| --- | --- | --- |
| Lag | Time(s) | Accuracy |
| 1 | 374 | 60.00% |
| 2 | 375 | 58.97% |
| 3 | 371 | 76.32% |
| 4 | 372 | 62.16% |
| 5 | 370 | 61.11% |

根據上述最優參數的設定後，增加迭代次數，可以發現在本研究探討台積電新聞的結果中是沒有明顯的成效，由表 4.3表示原先的設定的100次迭代已經達到很好的結果。

表 4.3使用本研究詞庫和不同迭代次數分析台積電新聞

|  |  |  |
| --- | --- | --- |
| Epoch | Time(s) | Accuracy |
| 100 | 382 | 76.32% |
| 200 | 762 | 60.53% |
| 300 | 1151 | 73.68% |

#### 鴻海

使用本研究詞庫分析鴻海新聞的LSTM架構中，LSTM Units使用256個神經元、Activation使用relu、Sequential Dropout設置為0、Recurrent Dropout設置為0、Loss Function採用MSE、Learning Rate設置為0.001能夠有效的學習鴻海新聞文章數量特徵，Dropout Layer是放置於LSTM隱藏層至Dense層之間，在鴻海新聞文章分析設置為0效果最佳，經過其他優化器比較後，決定使用adam演算法優化較為適合。上述設定由表 4.4表示，其參數皆為本研究經過Trial and error後獲得各個參數最佳值，Lag的表現中，由表 4.5可看出延遲4天所預測的精準度最高，由此可知鴻海在新聞分析中，Lag設為4天最為適當。

表 4.4使用本研究詞庫分析鴻海新聞所使用的參數

|  |  |
| --- | --- |
| LSTM Units | 256 |
| Activation | relu |
| Sequential Dropout | 0 |
| Recurrent Dropout | 0 |
| Loss Function | MSE |
| Learning Rate | 0.001 |
| Dropout Layer | 0 |
| Optimizers | adam |
| Epoch | 100 |

表 4.5使用本研究詞庫和不同的Lag分析鴻海新聞

|  |  |  |
| --- | --- | --- |
| Lag | Time(s) | Accuracy |
| 1 | 374 | 71.05% |
| 2 | 375 | 67.57% |
| 3 | 371 | 75.00% |
| 4 | 372 | 77.14% |
| 5 | 370 | 70.59% |

根據最優參數的設定後，增加迭代次數，可以發現在本研究探討鴻海新聞在本次迭代實驗中具有明顯的成效，由表 4.6可以看出100次迭代並不是最好的結果，200次迭代雖然花費較久的時間，不過可以得到較好的成果，至於300次迭代相較之下花費最久的時間，卻不是最好的結果，對於本次迭代實驗使用200次迭代為最優參數設定。

表 4.6使用本研究詞庫和不同迭代次數分析鴻海新聞

|  |  |  |
| --- | --- | --- |
| Epoch | Time(s) | Accuracy |
| 100 | 388 | 77.14% |
| 200 | 782 | 80.00% |
| 300 | 1227 | 80.00% |

#### 台塑化

使用本研究詞庫分析台塑化新聞的LSTM架構中，LSTM Units設定為128、Activation使用relu、Sequential Dropout設置為0.05、Recurrent Dropout設為0.05、Loss Function採用MSE評估實際值與預測值的誤差並修正，Learning Rate設置為0.0001能夠有效的學習台塑化新聞文章數量特徵，Dropout Layer在台塑化新聞文章分析設置為0.1效果最佳，比較其他優化器後，使用Nadam優化器分析台塑化新聞文章。上述設定由表 4.7表示，其參數皆為本研究經過Trial and error後獲得各個參數最佳值，Lag的表現中，由表 4.8可看出延遲3天所預測的精準度最高，由此可知台塑化在新聞分析中，Lag設為3天最為適當。

表 4.7使用本研究詞庫分析台塑化新聞所使用的參數

|  |  |
| --- | --- |
| LSTM Units | 128 |
| Activation | relu |
| Sequential Dropout | 0.05 |
| Recurrent Dropout | 0.05 |
| Loss Function | MSE |
| Learning Rate | 0.0001 |
| Dropout Layer | 0.1 |
| Optimizers | Nadam |
| Epoch | 100 |

表 4.8使用本研究詞庫和不同的Lag分析台塑化新聞

|  |  |  |
| --- | --- | --- |
| Lag | Time(s) | Accuracy |
| 1 | 112 | 51.00% |
| 2 | 112 | 48.39% |
| 3 | 113 | 48.28% |
| 4 | 113 | 50.00% |
| 5 | 114 | 48.15% |

根據最優參數的設定後，增加迭代次數，可以發現在本研究探討台塑化新聞在本次迭代實驗中具有明顯的成效，由表 4.9可以看出100次迭代並不是最好的結果，其結果比不上前面兩檔股票的精準度，探討訓練的過程中可以發現，迭代100次結束後，訓練特徵並沒有很好的被模型學習，增加迭代次數發現明顯提升精準度，若在提升迭代次數至300次為最佳結果若再增加迭代次數將會過度學習，導致測試結果不理想，迭代實驗對於台塑化新聞的情況下，設置為300次迭代為最優參數設定。

表 4.9使用本研究詞庫和不同迭代次數分析台塑化新聞

|  |  |  |
| --- | --- | --- |
| Epoch | Time(s) | Accuracy |
| 100 | 124 | 51.00% |
| 200 | 249 | 61.29% |
| 300 | 371 | 77.42% |

#### 個股(三類股)小結

對於個股的花費時間來說，在相同迭代的情況下，可以發現台積電與鴻海需要花費較久的時間，而台塑化花費的時間較少，其原因經過模型架構與參數設定探討後發現，LSTM Unit數量影響絕大多數的計算時間，增加個數會導致神經網路在調整、修正權重與參數所花費的時間以指數成長，由表 4.10表示。

表 4.10使用本研究詞庫分析個股新聞相同迭代次數訓練時間

|  |  |  |
| --- | --- | --- |
| 股票名稱 | Epoch | Time(s) |
| 台積電 | 100 | 382s |
| 鴻海 | 100 | 388s |
| 台塑化 | 100 | 124s |

在Lag的最佳設定中可以發現，台積電在Lag設定為3天的情況下表現最好，而鴻海與台塑化分別為4天、1天表現最優，由表 4.11我們可以了解每一檔股票對於新聞反應的天數都不相同，以個股分別實驗Lag的情況下，不難發現，其實Lag對於結果具有影響力，隨著Lag的變化訓練的資料也會因此有所改變，因此可知兩者確實具有強大的影響力。

台積電在迭代次數的實驗中，100次迭代已經可以達到最高的水準，若將次數往上提升，反而會降低準確度，分析鴻海時，在100次迭代時相較之下已經有一定水準的準確度，觀察訓練過程中發現，訓練特徵並沒有達到水準，為了學習更多訓練集的特徵，依然實作迭代次數實驗，在結果可以看出鴻海在300次迭代的結果已經沒有提升，因此200次迭代為最優次數，至於台塑化隨著迭代次數的增加而提升精準度，300次迭代為極限次數，若再往上提升迭代次數則會產生過度擬和的問題導致測試結果下滑

，其個股迭代次數結果比較由表 4.11所示。

根據實驗中三檔股票的參數設定後，比較最佳的結果可以發現，鴻海的最佳準確度較高，探討上述參數可以發現Dropout的設定有所不同，由Dropout可以看出訓練集與測試集之間的差異，本研究Dropout分為三個，Sequential Dropout、Recurrent Dropout、Dropout Layer，僅位置不同，功能皆相同，台積電於Sequential Dropout設置0.1，其餘皆沒有設置、鴻海在三者Dropout位置中，皆不需要設置，台塑化依序設置為0.05、0.05、0.1，訓練期間可以發現訓練集與測試集的特徵較台積電與台塑化來的相似，台塑化新聞文章中，需要更高的機率遮罩訓練集特徵並學習，才能將測試結果提升，其結果由表 4.11表示。

表 4.11使用本研究詞庫分析個股新聞最佳結果

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 股票名稱 | Lag | Epoch | Sequential Dropout | Recurrent Dropout | Dropout Layer | Accuracy |
| 台積電 | 3 | 100 | 0.1 | 0 | 0 | 76.32% |
| 鴻海 | 4 | 200 | 0 | 0 | 0 | 80.00% |
| 台塑化 | 1 | 300 | 0.05 | 0.05 | 0.1 | 77.42% |

### 台灣50預測大盤指數



#### 本節實驗目的

* 蒐集台灣市值前50檔股票相關新聞做語意分析是否可以精準預測大盤指數？
* 新聞反應的時間對於大盤指數是否具有影響力？
* 新聞反應至大盤指數最佳天數為何？

#### 台灣50預測

使用本研究詞庫分析台灣50相關新聞在LSTM架構中LSTM Units使用256個神經元、Activation使用relu、Sequential Dropout設置為0.1、Recurrent Dropout設置為0.2、Loss Function採用MSE評估修正，Learning Rate設置為0.001能夠有效的學習台灣50相關新聞文章數量特徵，Dropout Layer在台灣50相關新聞文章分析中將參數設置為0.3效果最佳，經過其他優化器比較後，決定使用adam演算法優化較為適合。上述設定由表 4.12表示，其參數皆為本研究經過Trial and error後獲得各個參數最佳值，Lag的表現中，由表 4.13可明顯看出延遲2天所預測的精準度能夠表現比其他天都要來的好，由此判定台灣50相關新聞分析中，Lag設為2天最為適當。

表 4.12使用本研究詞庫分析台灣50相關新聞所使用的參數

|  |  |
| --- | --- |
| LSTM Units | 256 |
| Activation | relu |
| Sequential Dropout | 0.1 |
| Recurrent Dropout | 0.2 |
| Loss Function | MSE |
| Learning Rate | 0.001 |
| Dropout Layer | 0.3 |
| Optimizers | adam |
| Epoch | 100 |

表 4.13使用本研究詞庫和不同的Lag分析台灣50相關新聞

|  |  |  |
| --- | --- | --- |
| Lag | Time(s) | Accuracy |
| 1 | 248 | 56.82% |
| 2 | 249 | 72.09% |
| 3 | 248 | 57.14% |
| 4 | 246 | 58.54% |
| 5 | 250 | 55.00% |

根據上述最優參數的設定後，增加迭代次數至200次時，訓練精準度大幅提升，模型學習到部分訓練集特徵，測試精準度明顯提升從100次迭代的48.84%提升至72.09%，增加至300次迭代時，訓練精準度更完整的學習，導致測試集特徵部分遺失，測試結果下降至69.77%，由表 4.14可以表示上述精準度變化，並得知200次迭代對於本研究詞庫分析台灣50資料集能夠達到最佳的結果。

表 4.14使用本研究詞庫和不同迭代次數分析台灣50相關新聞

|  |  |  |
| --- | --- | --- |
| Epoch | Time(s) | Accuracy |
| 100 | 273 | 48.84% |
| 200 | 276 | 72.09% |
| 300 | 404 | 69.77% |

#### 台灣50預測大盤指數小結

從個股與大盤的實驗中可以發現個股與股價之間的Lag有所不同，台灣50的Lag與個股也沒有相關，全都取決於資料集特徵，台灣50對於大盤指數的Lag為2天的精準度相較之下也在一定的水準，台塑化在Lag的比較中雖然在水準之下，由圖 4.1所示，增加跌代後仍然能夠擁有很好的結果，使用本研究的模型與實驗流程對於個股能夠擁有較好的結果，在訓練期間，台灣50平均需要的迭代次數較多，才會開始學習到部分訓練集特徵，測試集才會有所變化，其次是台塑化，台積電、鴻海不需要太多迭代次數就能夠得到最好的結果，台灣50在300次迭代時過度擬和訓練集造成測試精準度下滑，如表 4.15表示最佳參數與迭代次數。

圖 4.1 本研究詞庫分析個股、大盤股價對於新聞的Lag

表 4.15使用本研究詞庫分析個股與台灣50相關新聞最佳結果

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 股票名稱 | Lag | Epoch | Sequential Dropout | Recurrent Dropout | Dropout Layer | Accuracy |
| 台積電 | 3 | 100 | 0.1 | 0 | 0 | 76.32% |
| 鴻海 | 4 | 200 | 0 | 0 | 0 | 80.00% |
| 台塑化 | 1 | 300 | 0.05 | 0.05 | 0.1 | 77.42% |
| 台灣50 | 2 | 200 | 0.1 | 0.2 | 0.3 | 72.09% |

### 台灣大學公開的極性詞庫(NTUSD)



#### 本節實驗目的

* 採用NTUSD詞庫是否能夠使用本研究LSTM模型預測未來趨勢達一定水準？
* NTUSD詞庫在三檔股票和大盤指數中，新聞最佳Lag設定為何？
* 本研究與NTUSD詞庫在此模型設計中何者較優？

#### 台積電

NTUSD詞庫分析台積電新聞在LSTM架構中LSTM Units使用256個神經元、Activation使用relu、Sequential Dropout設置為0.1、Recurrent Dropout設置為0.1、Loss Function採用MSE評估實際值與預測值的誤差並修正，Learning Rate設置為0.0005能夠有效的學習台積電新聞文章數量特徵，Dropout Layer在台積電新聞文章分析設置為0.3效果最佳，經過其他優化器比較後，決定使用Nadam演算法優化較為適合。上述設定由表 4.16表示，其參數皆為本研究經過Trial and error後獲得各個參數最佳值，Lag的表現中，由表 4.17可看出延遲4天所預測的精準度最高，由此可知台積電新聞分析中，Lag設為4天最為適當。

表 4.16 NTUSD詞庫分析台積電新聞所使用的參數

|  |  |
| --- | --- |
| LSTM Units | 256 |
| Activation | relu |
| Sequential Dropout | 0.1 |
| Recurrent Dropout | 0.1 |
| Loss Function | MSE |
| Learning Rate | 0.0005 |
| Dropout Layer | 0.3 |
| Optimizers | Nadam |
| Epoch | 100 |

表 4.17 NTUSD詞庫和不同的Lag分析台積電新聞

|  |  |  |
| --- | --- | --- |
| Lag | Time(s) | Accuracy |
| 1 | 385 | 60.00% |
| 2 | 382 | 58.97% |
| 3 | 377 | 60.53% |
| 4 | 388 | 67.57% |
| 5 | 381 | 66.67% |

根據上述最優參數的設定後，增加迭代次數，可以發現在本研究探討台積電新聞的結果中是沒有明顯的成效，由表 4.18表示原先的設定的100次迭代已經達到很好的結果。

表 4.18 NTUSD詞庫和不同迭代次數分析台積電新聞

|  |  |  |
| --- | --- | --- |
| Epoch | Time(s) | Accuracy |
| 100 | 383 | 67.57% |
| 200 | 767 | 58.97% |
| 300 | 1145 | 61.54% |

#### 鴻海

使用本研究詞庫分析鴻海新聞在LSTM架構中LSTM Units使用128個神經元、Activation使用relu、Sequential Dropout設置為0.1、Recurrent Dropout與Dropout的功用相同，數值設置為0、Loss Function採用MSE評估實際值與預測值的誤差並修正，Learning Rate設置為0.0005能夠有效的學習鴻海新聞文章數量特徵，Dropout Layer在鴻海新聞文章分析設置為0.3效果最佳，經過其他優化器比較後，決定使用Nadam演算法優化較為適合。上述設定由表 4.19表示，其參數皆為本研究經過Trial and error後獲得各個參數最佳值，Lag的表現中，由表 4.20可看出延遲1天所預測的精準度最高，由此可知鴻海新聞分析中，Lag設為1天最為適當。

表 4.19 NTUSD詞庫分析鴻海新聞所使用的參數

|  |  |
| --- | --- |
| LSTM Units | 128 |
| Activation | relu |
| Sequential Dropout | 0.1 |
| Recurrent Dropout | 0 |
| Loss Function | MSE |
| Learning Rate | 0.0005 |
| Dropout Layer | 0.3 |
| Optimizers | Nadam |
| Epoch | 100 |

表 4.20 NTUSD詞庫和不同的Lag分析鴻海新聞

|  |  |  |
| --- | --- | --- |
| Lag | Time(s) | Accuracy |
| 1 | 134 | 76.32% |
| 2 | 135 | 70.27% |
| 3 | 136 | 66.67% |
| 4 | 136 | 65.71% |
| 5 | 137 | 61.77% |

根據上述最優參數的設定後，增加迭代次數，原先設定的100次迭代已經是最優結果，越高的迭代反而會造成過度學習訓練集特徵，遺失原先記憶的測試集特徵，因此在本研究探討鴻海新聞的結果中，增加迭代次數是沒有效果的，由表 4.21表示原先的設定的100次迭代已經達到很好的結果。

表 4.21 NTUSD詞庫和不同迭代次數分析鴻海新聞

|  |  |  |
| --- | --- | --- |
| Epoch | Time(s) | Accuracy |
| 100 | 383 | 76.32% |
| 200 | 767 | 67.57% |
| 300 | 1145 | 64.87% |

#### 台塑化

使用本研究詞庫分析台塑化新聞在LSTM架構中LSTM Units使用64個神經元、Activation使用relu、Sequential Dropout設置為0.1、Recurrent Dropout與Dropout的功用相同，數值設置為0.1、Loss Function採用MSE評估實際值與預測值的誤差並修正，Learning Rate設置為0.001能夠有效的學習台塑化新聞文章數量特徵，Dropout Layer在台塑化新聞文章分析設置為0.3效果最佳，經過其他優化器比較後，決定使用Nadam演算法優化較為適合。上述設定由表 4.22表示，其參數皆為本研究經過Trial and error後獲得各個參數最佳值，Lag的表現中，由表 4.23可看出延遲3天所預測的精準度最高，由此可知台塑化新聞分析中，Lag設為3天最為適當。

表 4.22 NTUSD詞庫分析台塑化新聞所使用的參數

|  |  |
| --- | --- |
| LSTM Units | 64 |
| Activation | relu |
| Sequential Dropout | 0.1 |
| Recurrent Dropout | 0.1 |
| Loss Function | MSE |
| Learning Rate | 0.001 |
| Dropout Layer | 0.3 |
| Optimizers | Nadam |
| Epoch | 100 |

表 4.23 NTUSD詞庫和不同的Lag分析台塑化新聞

|  |  |  |
| --- | --- | --- |
| Lag | Time(s) | Accuracy |
| 1 | 90 | 70.97% |
| 2 | 91 | 56.67% |
| 3 | 93 | 75.86% |
| 4 | 92 | 60.71% |
| 5 | 92 | 55.56% |

根據上述最優參數的設定後，增加迭代次數，原先設定的100次迭代已經是最優結果，越高的迭代反而會造成過度學習訓練集特徵，遺失原先記憶的測試集特徵，模型學到越深層的訓練集特徵，反而會造成測試集誤差提升，由此可知，訓練集深層的特徵與測試集的特徵不相似，由200次以上的迭代發現此現象，由表 4.24表示原先的設定的100次迭代已經達到很好的結果。

表 4.24 NTUSD詞庫和不同迭代次數分析台塑化新聞

|  |  |  |
| --- | --- | --- |
| Epoch | Time(s) | Accuracy |
| 100 | 95 | 75.86% |
| 200 | 192 | 60.00% |
| 300 | 299 | 63.33% |

#### 台灣50預測大盤指數

使用本研究詞庫分析台灣50相關新聞在LSTM架構中LSTM Units使用64個神經元、Activation使用sigmoid、Sequential Dropout設置為0、Recurrent Dropout與Dropout的功用相同，數值設置為0、Loss Function採用MSE評估實際值與預測值的誤差並修正，Learning Rate設置為0.001能夠有效的學習台灣50相關新聞文章數量特徵，Dropout Layer在台灣50相關新聞文章分析中將參數設置為0.1效果最佳，經過其他優化器比較後，決定使用adam演算法優化較為適合。上述設定由表 4.25表示，其參數皆為本研究經過Trial and error後獲得各個參數最佳值，Lag的表現中，由表 4.26可以看出第1天與第5天的精準度都很高，其中第1天精準度略微高於第5天，由此可知台灣50相關新聞分析中，Lag設為1天最為適當。

表 4.25 NTUSD詞庫分析台灣50相關新聞所使用的參數

|  |  |
| --- | --- |
| LSTM Units | 64 |
| Activation | sigmoid |
| Sequential Dropout | 0 |
| Recurrent Dropout | 0 |
| Loss Function | MSE |
| Learning Rate | 0.001 |
| Dropout Layer | 0.1 |
| Optimizers | adam |
| Epoch | 100 |

表 4.26 NTUSD詞庫和不同的Lag分析台灣50相關新聞

|  |  |  |
| --- | --- | --- |
| Lag | Time(s) | Accuracy |
| 1 | 248 | 70.46% |
| 2 | 249 | 67.44% |
| 3 | 248 | 61.91% |
| 4 | 246 | 60.98% |
| 5 | 250 | 70.00% |

根據上述最優參數的設定後，以表 4.26中最佳結果為基底，增加迭代次數實驗，表 4.27可以發現，原先設定的100次迭代並不是最佳的結果，探討訓練期間精準度變化後可以發現，越多次的迭代可以讓模型學習更多的訓練集特徵，當迭代次數達到300次時，訓練精準度表示模型已經相當完整的學習訓練集特徵，對於測試集也有一定程度的提升。

表 4.27 NTUSD詞庫和不同迭代次數分析台灣50相關新聞

|  |  |  |
| --- | --- | --- |
| Epoch | Time(s) | Accuracy |
| 100 | 273 | 70.00% |
| 200 | 276 | 70.46% |
| 300 | 404 | 71.43% |

* + 1. **台灣大學公開的極性詞庫(NTUSD)詞庫小結**

經過最佳參數設置後，NTUSD詞庫用於本研究架構分析新聞文章Lag可以發現每一檔股票對於新聞反應的天數都不相同，由表 4.28表示個股最佳結果的Lag，台積電相關新聞反應至股價的時間為3天的表現結果最好，鴻海最佳Lag為4天，台塑化、台灣50需要1天的Lag，不同的Lag牽扯到訓練資料，觀看測試結果可以發現確實具有影響力。

根據實驗中三檔股票與大盤的參數設定後，比較最佳的結果可以發現，鴻海的最佳準確度較高，探討上述參數可以發現Dropout的設定有所不同，由Dropout可以看出訓練集與測試集之間的差異，Sequential Dropout、Recurrent Dropout、Dropout Layer，台積電、台塑化設置相同的參數，於Sequential Dropout設置0.1、Recurrent Dropout設置0.1、Dropout Layer設置0.3，相較之下，需要較高的機率遮罩才能降低過度擬和提升測試集精準度，鴻海在三者Dropout位置中，Sequential Dropout設置0.1，而台灣50則是設置0.1 Dropout Layer的位置，鴻海與台灣50在另外兩個位置皆不需要設置就能得到最佳結果，其結果由表 4.28表示。

表 4.28 NTUSD詞庫分析個股與大盤最佳結果

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 股票名稱 | Lag | Epoch | Sequential Dropout | Recurrent Dropout | Dropout Layer | Accuracy |
| 台積電 | 4 | 100 | 0.1 | 0.1 | 0.3 | 67.57% |
| 鴻海 | 1 | 100 | 0.1 | 0 | 0.3 | 76.32% |
| 台塑化 | 3 | 100 | 0.1 | 0.1 | 0.3 | 75.86% |
| 台灣50 | 1 | 300 | 0 | 0 | 0.1 | 71.43% |

### 實驗結果討論

圖 4.2 本研究與台灣大學詞庫分析個股新聞文章最佳結果比較

詞庫比較中，台積電使用本研究自定義詞庫最佳結果為76.32%，在NTUSD詞庫獲得的最佳結果為67.57%，而在鴻海使用本研究自定義詞庫最佳結果為80.00%，在NTUSD詞庫獲得的最佳結果為76.32%，在台塑化使用本研究自定義詞庫與NTUSD詞庫獲得的最佳結果為依序為77.42%、75.86%，兩者詞庫比較可以發現至少有1.56%的精準度差距，使用台灣50預測大盤指數時本研究自定義詞庫最佳結果為72.09%，在NTUSD詞庫獲得的最佳結果為71.43%，其差距僅有0.66%，由此可知，個股預測的情況使用本研究詞庫效果明顯較佳，使用台灣50預測大盤指數的情況兩者詞庫僅不到1%的差距。

在類別比較可以發現，個股不論是使用本研究詞庫或NTUSD詞庫相較於台灣50來說，都能夠獲得較高的精準度，其大盤指數預測採用台灣50相關新聞文章預測，台灣50是50檔股票，以研究參數來看，每一個股票都擁有自己的延遲時間，能夠有效預測的參數也有所不同，將所有新聞文章結合預測，同一組參數會導致某幾檔結果較優、某幾檔結果較差，以原始資料觀點來說，每個股票能夠保留的歷史資料與熱門報導程度有所不同，若不考慮近期熱門因素，針對市值佔大盤比率較低的個股所擁有的相關新聞語意強度、文章資料量也相對較低，對於統計後的變化量相對也沒有那麼明顯的特徵，模型在學習時針對不明顯的特徵也很難學習訓練集特徵，上述兩種觀點說明台灣50預測大盤平均結果輸給單一個股預測之情形，由圖 4.2表示。

表 4.29使用本研究詞庫與NTUSD詞庫最佳結果比較表

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 本研究詞庫 | | | | | | | |
| 股票名稱 | Lag | Epoch | Sequential Dropout | Recurrent Dropout | Dropout Layer | Accuracy | |
| 台積電 | 3 | 100 | 0.1 | 0 | 0 | 76.32% | |
| 鴻海 | 4 | 200 | 0 | 0 | 0 | 80.00% | |
| 台塑化 | 1 | 300 | 0.05 | 0.05 | 0.1 | 77.42% | |
| 台灣50 | 2 | 200 | 0.1 | 0.2 | 0.3 | 72.09% | |
| NTUSD詞庫 | | | | | | |
| 股票名稱 | Lag | Epoch | Sequential Dropout | Recurrent Dropout | Dropout Layer | Accuracy | |
| 台積電 | 4 | 100 | 0.1 | 0.1 | 0.3 | 67.57% | |
| 鴻海 | 1 | 100 | 0.1 | 0 | 0.3 | 76.32% | |
| 台塑化 | 3 | 100 | 0.1 | 0.1 | 0.3 | 75.86% | |
| 台灣50 | 1 | 300 | 0 | 0 | 0.1 | 71.43% | |

從表 4.29明顯看出兩種詞庫對於新聞分析中，反應天數、迭代次數、Dropout數值與位置的不同，以平均迭代次數來說NTUSD得到最佳結果的所需要的迭代次數較少，本研究詞庫所需要的迭代次數較多，訓練特徵較難被模型學習，一旦被模型學習，即可提高測試精準度，所需要的Dropout數量與位置也較少，不需要遮蔽太多訓練集特徵就能夠成功預測測試集。

## 結論

### 研究貢獻

本研究自行定義在金融相關領域極性詞庫，並在預測個股股價與大盤指數都達到一定水準，一般詞庫受限於主題、專業領域、辨識率不高等等因素，在此使用客觀的方式定義極性詞庫，不受限於某專業或特殊領域主題，能夠根據資料特徵、資料表示的詞彙變化，而本研究以個股、大盤為目標欄位，財經新聞作為訓練資料並定義該領域相關極性詞彙，在結果呈現中表示也能夠有效預測股價未來走勢。

詞庫的比較中與具有泛用性的NTUSD詞庫進行比較，在本研究架構中自建詞庫擁有較好的預測結果，其中NTUSD詞庫預測分析新聞文章語意後，經過LSTM模型架構預測個股與大盤指數，原先NTUSD詞庫並不是使用新聞文章訓練極性詞彙，由此證實NTUSD詞庫具有一定程度的泛用性，若使用本研究定義詞彙方式，能夠更精準預測股價、指數未來趨勢，對於詞彙的極性判定方式，採用傳統權重計算方式，為了提升各個領域的相容性，不再是侷限於某特定領域、專業領域、主題的詞庫。

不管是個股、大盤延遲程度如何，根據上述實驗，都能夠證實在台灣個股和大盤指數都會受到相關新聞文章的極性影響，由此可知，新聞文章和每日數據進行預測確實是與個股漲跌具有相關性，也證實投資者觀看這些數據找出規律、猜測、訂定投資策略確實相當可靠。在機器學習的角度下，模型透過相同的方式訓練並且了解這些數據的特徵點，期望它能夠與人擁有相同的判斷能力，以往的經驗來說，發現某特定的時間點出現相同的趨勢、出現某趨勢而下一個趨勢又是如何等等類似的規律性能夠被挖掘，不需要再花費人們寶貴的時間做繁瑣、重複的事情，且在市面上有些財經專家講解技術分析，也是透過經驗累積而成，且將經驗融會貫通，再將此經驗分享給需要的人，而本次實驗模仿相同的模式學習，經過時間與歷史資料的訓練後，盼望能與專家一樣，能夠擁有預測未來趨勢的能力。

閱讀文章需要花費大量的時間，而這些新聞也能夠對於投資者有所影響，網路上的新聞文章不考慮操弄的情況下，本實驗模擬投資者觀看新聞能夠了解未來走勢的想法，實驗證實觀看新聞能有一定程度判斷未來走勢的能力，文字探勘中依據的詞庫最為重要，本實驗中建置專屬於金融、股票新聞詞庫，這個詞庫能夠用於各個領域，在使用之前必須去除金融專有名詞、專業術語後就能夠廣泛運用，文本中有許多字詞是共通的，能夠根據每天漲跌的描述新聞決定詞彙的極性，若資料量夠大，就能夠找出下跌所描述的文字和上漲所描述的文字有所不同，並且加以定義權重，增加關鍵字詞影響力，也能夠去除不必要的文字和雜訊。

### 研究限制

當前數據蒐集種類有限，根據隱私、個資法規相關規定蒐集資料，細節資料無法深入獲取，因此僅能探討一般使用者能獲取的資訊範圍，像是財報、新聞、每日數據、交易量等等公開資料，這些公開資料可能會經過處理後公開，此決策可能牽扯到公司政策、個人因素、會計人員等等，相關數據過於困難收集與探勘，因此本研究僅探討一般使用者能獲得的資訊為實驗數據，在特徵預測方面，能夠獲得的數據很多，本研究根據盤後資訊訓練分析，以一天為一個粒度，並無及時預測，且根據過往資料進行預測，當訓練集無法有效被模型學習，誤差往往會過度訓練某一天，導致過適化的問題。

現今已經出現文章產生器，不論他寫的是不是真的有意涵，對於不知情的使用者甚至會信以為真，真正有深度的投資者會收集相關資料，辨別事件真假，但這其中佔少數，大多的使用者皆相信文章所描述的內容，有些特例公司對外釋放的造假消息，新聞大肆傳播下，許多投資客在此時作出錯誤的判斷，擁有內線消息者在這期間快速脫手所擁有的相關資產、股票，上述案例無法在常態分析中深入探討，外部投資客不可能知道的消息，學習模型就別說能夠預測了，連專家經過經驗判讀能夠準確地都占少數，且學習案例過少，就算過去歷史資料擁有發生事件前的相關新聞，本研究並沒有深入探討內線消息對外是否有刻意散播造假消息等因素。

台灣50以2016年統計市值排名前50作為評估大盤的標準，在實驗中台灣50每一隻股票對於大盤的影響程度相同，未考量每一隻股票對於大盤的影響權重，由於本實驗橫跨兩年期間，2017年市值排名有所變動，因變動不大，不考量此因素。

### 未來研究方向

這樣的結果在實務上是沒有辦法穩定的在投資，入場的時機點、出場的時機點、買進賣出多少股數等等動作，這些情況的發生無法即時的模型沒有辦法應對，若能夠加入更多可能會影響股價的因素、變數、參數等等，讓輸入數據對於股價能有更多的影響力，進而提升個股與大盤的精準度，達到更高的水準後，能夠進行投資報酬率的分析，甚至是即時的預測分析，更精確的讓分析者能夠評估進出場時間點。投資者很容易受到新聞、政治因素影響，許多新聞文章無法判定是否為寫手所撰寫的，為了要影響選民、市民的心態，不論是對哪一個政黨可能都隱藏著這樣的因素，若能夠加入政治新聞進行分析預測對於股價影響將會是一個很重要的因素之一。

由實驗我們可以知道市值排序前50名能夠預測大盤指數達一定水準，評估影響力採用市值排行的方式，選擇2016前50名分析預測2016年初至2017年底，對於2017年數據來說，2016年排名前50名的個股可能因此有所變動，若在預測大盤指數時，台灣50以動態時間點的方式認定影響大盤前50大的個股作為分析標準，若大盤訓練預測2010年至2018年，擷取的數據應以2010年個股市值排行、2011個股市值排行以此類推至2018年，如此一來，每一年對於大盤指數都會有一定水準的影響力，也能夠改善對於大盤預測的能力。

## **參考文獻**

1. 江旻緯. (2014) 以改良式支持向量機建立股票股價漲跌預測模型. 國立高雄應用科技大學資訊管理系碩士論文
2. 李春淋. (2010). 個股新聞對股價影響之研究-以台股為例，輔仁大學應用統計學系碩士論文。
3. 喻欣凱. (2008). 運用支援向量機與文字探勘於股價漲跌趨勢之預測. 輔仁大學資訊管理學系碩士論文, 新北市.
4. 新華字典辭典 (2013). <http://xh.5156edu.com/page/z8338m5798j18550.html>. Accessed 2 June 2018
5. 謝政翰. (2017) 應用機器學習演算法與模糊推論於股價漲跌預測模式建構之研究. 國立聯合大學資訊管理學系碩士論文
6. Araque, O., Corcuera-Platas, I., Sanchez-Rada, J. F., & Iglesias, C. A. (2017). Enhancing deep learning sentiment analysis with ensemble techniques in social applications. Expert Systems with Applications, 77, 236-246.
7. Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. Journal of computational science, 2(1), 1-8.
8. Church, K. W., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. Computational linguistics, 16(1), 22-29.
9. Chan, S. W., & Chong, M. W. (2017). Sentiment analysis in financial texts. Decision Support Systems, 94, 53-64.
10. Cao, L., Xu, L., & Goodman, E. D. (2018). A neighbor-based learning particle swarm optimizer with short-term and long-term memory for dynamic optimization problems. Information Sciences, 453, 463-485.
11. Deng, S., Sinha, A. P., & Zhao, H. (2017). Adapting sentiment lexicons to domain-specific social media texts. Decision Support Systems, 94, 65-76.
12. Fu, X., Liu, W., Xu, Y., & Cui, L. (2017). Combine HowNet lexicon to train phrase recursive autoencoder for sentence-level sentiment analysis. Neurocomputing, 241, 18-27.
13. García, A. (2016). Global financial indices and twitter sentiment: A random matrix theory approach. Physica A: Statistical Mechanics and its Applications, 461, 509-522.
14. Gadia, K., & Bhowmick, K. (2015). Parallel text mining in multicore systems using FP-Tree algorithm. Procedia Computer Science, 45, 111-117.
15. Giatsoglou, M., Vozalis, M. G., Diamantaras, K., Vakali, A., Sarigiannidis, G., & Chatzisavvas, K. C. (2017). Sentiment analysis leveraging emotions and word embeddings. Expert Systems with Applications, 69, 214-224.
16. Guo, K., Sun, Y., & Qian, X. (2017). Can investor sentiment be used to predict the stock price? Dynamic analysis based on China stock market. Physica A: Statistical Mechanics and its Applications, 469, 390-396.
17. Hung, C. (2017). Word of mouth quality classification based on contextual sentiment lexicons. Information Processing & Management, 53(4), 751-763.
18. Huang, D., Jiang, Z., Zou, L., & Li, L. (2017). Drug–drug interaction extraction from biomedical literature using support vector machine and long short term memory networks. Information sciences, 415, 100-109.
19. Jaccard, P. (1902). Distribution comparée de la flore alpine dans quelques régions des Alpes occidentales et orientales. Bulletin de la Murithienne, (31), 81-92.
20. fxsjy(2017). https://github.com/fxsjy/jieba, Accessed 2 June 2018.
21. Li, N., & Wu, D. D. (2010). Using text mining and sentiment analysis for online forums hotspot detection and forecast. Decision support systems, 48(2), 354-368.
22. Lin, D., Li, L., Cao, D., Lv, Y., & Ke, X. (2018). Multi-modality weakly labeled sentiment learning based on Explicit Emotion Signal for Chinese microblog. Neurocomputing, 272, 258-269.
23. Musoll, E., Lang, T., & Cortadella, J. (1998). Working-zone encoding for reducing the energy in microprocessor address buses. IEEE Transactions on Very Large Scale Integration (VLSI) Systems, 6(4), 568-572.
24. Ma, W. Y., & Chen, K. J. (2003, July). Introduction to CKIP Chinese word segmentation system for the first international Chinese Word Segmentation Bakeoff. In Proceedings of the second SIGHAN workshop on Chinese language processing-Volume 17 (pp. 168-171). Association for Computational Linguistics..
25. Ma, W. Y., & McKeown, K. (2012). Detecting and correcting syntactic errors in machine translation using feature-based lexicalized tree adjoining grammars. International Journal of Computational Linguistics & Chinese Language Processing, Volume 17, Number 4, December 2012-Special Issue on Selected Papers from ROCLING XXIV, 17(4).
26. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems (pp. 3111-3119).
27. Meire, M., Ballings, M., & Van den Poel, D. (2016). The added value of auxiliary data in sentiment analysis of Facebook posts. Decision Support Systems, 89, 98-112.
28. Ortigosa, A., Martín, J. M., & Carro, R. M. (2014). Sentiment analysis in Facebook and its application to e-learning. Computers in human behavior, 31, 527-541.
29. Pandey, A. C., Rajpoot, D. S., & Saraswat, M. (2017). Twitter sentiment analysis using hybrid cuckoo search method. Information Processing & Management, 53(4), 764-779.
30. Poria, S., Peng, H., Hussain, A., Howard, N., & Cambria, E. (2017). Ensemble application of convolutional neural networks and multiple kernel learning for multimodal sentiment analysis. Neurocomputing, 261, 217-230.
31. Romaine, S., & Algeo, J. (1992). The Cambridge history of the English language (Vol. 3). Cambridge University Press.
32. Robertson, S. (2004). Understanding inverse document frequency: on theoretical arguments for IDF. Journal of documentation, 60(5), 503-520.
33. Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. Journal of documentation, 28(1), 11-21.
34. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research, 15(1), 1929-1958.
35. Wu, H. C., Luk, R. W. P., Wong, K. F., & Kwok, K. L. (2008). Interpreting tf-idf term weights as making relevance decisions. ACM Transactions on Information Systems (TOIS), 26(3), 13.
36. Wang, H., Wu, J., Yuan, S., & Chen, J. (2016). On characterizing scale effect of Chinese mutual funds via text mining. Signal Processing, 124, 266-278.
37. Zhang, D., Xu, H., Su, Z., & Xu, Y. (2015). Chinese comments sentiment classification based on word2vec and SVMperf. Expert Systems with Applications, 42(4), 1857-1863.