應用事件驅動技術於股價漲跌預測

指導教授：曾意儒老師

組員：鄭凱元、蔡承運、張茜茜、何欣恬

分工：

鄭凱元 ： 蒐集資料與前處理、word embedding、分類與視覺化

蔡承運 ：event embedding、分類

張茜茜、何欣恬 ： 做簡報、找爬蟲的網站

一、背景及動機

股市的漲跌是我們都很難預料的，但有時候或許可以從一些生活中的眉角稍稍找到股票中的股性。而我們發現明顯影響美國股票漲跌有兩大推手，第一是美國總統川普、第二是特斯拉的執行長馬斯克，只要他們在 Twitter 上發布一些文章，隔天美國股市就會有明顯的變化，又由於今年新冠肺炎(COVID-19)疫情嚴重影響到美國的經濟，進而引響到股價的漲跌，因此我們想藉此分析一些有關 COVID-19 在新聞網上的文章和股票漲跌之間的關聯性，也探討文字影響股市這件事情的關聯性有多大。

二、資料來源

資料來源是來自 CNN 與 Drudge Report 兩個新聞網站，會選取這兩個網站的原因如下，第一條件先以政治立場來探討，由於不同新聞台的言論可能因著不同的政治立場會有偏頗，所以我們需要考量到各家新聞台的政治立場再做分析，因此我們選擇較為偏左派(民主黨)的 CNN 新聞網以及較為偏右派(共和黨)的 Drudge Report 新聞網。第二條件是以爬蟲的方便性來探討，因著各新聞網有不同的網站結構，所以我們篩選掉有些結構較複雜、較難爬取的網站，因此在一次爬取多頁和使用瀑布式爬蟲時，能夠較為快速且正確性高的抓取到欲分析的內容。第三條件是由於我們需要以天數作為分析的依據，所以我們所要找尋的網站需要有明確的新聞時間點，所以我們篩選掉有些日期顯示不完整的

網站。由於以此三種原因作為篩選網站的條件，因此我們以 CNN 和 Drudge Report 此二新聞網作為分析的資料來源。

三、股價預測方法

（一）研究架構

我們的專題有部分是參考 Ding et al.（2015）的研究，但有稍做修改。首先，我們先從兩家新聞媒體網站以爬蟲程式取得新聞的標題及內文，將取得的標題及內文輸入 Skip-Gram 演算法，訓練出 word embedding model。接著，將新聞標題數入 OpenIE 中萃取出主詞、動詞、受詞（O1, P, O2），將萃取出的主詞動詞受詞輸入 Word embedding model 中，經過平均計算後，輸出 word embedding（O1, P, O2 各一個 500 維的向量）。接著將得到的三個 500 維向量輸入訓練好的 event embedding model（Neural Tensor Network）中，最後輸出一個 50 維的向量表示萃取出的事件在多維空間中的位置。接著把 50 維向量的每一維（一個數字）當作一個 column，以羅吉斯回歸模型預測各個事件隔天的漲跌（1 or 0），再以 step AIC 驗算法挑選適合的模型。

（二）爬蟲方法

我們爬蟲總共使用了三個套件，分別是（１）rvest：利用裡面的函數讀取網頁網頁、抓取節點、擷取內容和擷取資料參數、（２）dplyr：利用 select 函數選取要分析的欄位、（３）RSelenium：在 R 中使用 selenium 自動操控瀏覽器，下載並擷取網頁中的資料。而爬蟲的邏輯以 CNN 為例是先到 CNN 的新聞網站裡，手動篩選出和 cornonavirus 相關的所有新聞，接著取得各天的新聞網址(CNN 的網址是由日期組成的)， 由於新聞網址是有規律性的，所以可以使用 paste0 再透過更改日期的區間抓取想要分析的網址，再丟進 RSelenium 產生的自動操控瀏覽器裡，並設定滾輪滾 30 下確保一天裡的所有新聞都有被載入出來後，就可以透過抓取到的網址連結到此篇新聞，再加以抓取其日期以及內容作為分析。

（三）Word Embedding

為了將一字串做計算處理，而有了 Word embedding 一個將文字轉成數值向量的概念，這個向量統稱詞向量；為了讓 neural tensor network 在 event embedding 前，有合適的詞向量做為模型輸入，所以我們先將爬蟲得到的新聞內容做資料清洗，使用 skip-gram(相比 CBOW 或 n-gram 得到的結果更好)一個根據上下文來做 embedding 的工具，來訓練 word embedding 的模型；接著用 stanford 提供的 openIE api 提取新聞標題的主詞、動詞、受詞形成一個字串型態的 triple，但此時因為是主、動、受詞的字串型態，所以再利用 embedding 模型將字串型態的 triple 轉成詞向量，並把他們輸入訓練好的 event embeding model 中。

（四）Event Embedding

我們參考 Ding et al.（2015）的研究，使用 Neural Tensor Network（NTN）作為研究的事件驅動模型，其優點為可以處理不同順序但同意義的事件物件。NTN 模型原先被應用於判斷關係的正確程度（Socher et al., 2013），其利用張量乘積的方式處理不同實體（entity）的關係，有效勝過其他傳統的模型，因此我們使用三層 NTN 作為將 word embedding 轉換成 event embedding 的模型。第一層中，我們將 O1、P 輸入模型中，產生 R1，第二層中，我們將 P、O2 輸入模型，產生 R2，第三層中，我們將 R1、R2 輸入模型產生 U（最後的 event embedding），其中第一層的公式如下：

$$R1 = f(O1TT1[1:k]P+WO1, P+b) \quad （1）$$

其中 T1 為一個 k 層的張量，W、b 為參數，f 為 tanh。第二層級第三層的公式皆與式（1）相同，只是將不同元素帶入公式中。模型的更新方法我們參考了一般 margin loss fuction 的做法、Socher et al.（2013）的方法以及 Ding et al.（2015）的方法。首先，我們先為每一個 event（E; O1, P, O2 的 word embedding）製作一個相對的被打亂的 event（Er）。Er 的製作方式為將每個 E 的 O1 隨機取代為其他 event 的 O1。接著，將 E、Er 輸入 NTN 中，各算出一個 50 維的向量，並用其計算其 margin loss，公式如下：

$$\text{margin loss}=\max(0, 1-f(Ei)-f(Eir)22) \quad （2）$$

其中 f(Ei)為 E 的 event embedding（NTN 的輸出），f(Eir)為 Er 的 event embedding。若是 margin loss 大於 0，則以反向誤差傳播法取得微分值（在此參考了 Socher et al.的張量反向傳播方式），並以 SGD 方法更新參數（Socher et al.中使用的是 AdaGrad，但在此為方便直觀調整不同參數的學習率，我們使用

的是 SGD 方法）。而若是 margin loss 大於或等於 0，則從所有 event 中取出現在的 E，並繼續計算下一個 E 及 Er 的 margin loss。

四、預測方法及結果

我們假設新聞中 title 的主詞動詞受詞對股票的漲跌有一定的影響關係，因此當 event embedding 產生出 50 維的向量後；我們的預測方法用新聞標題對股票的上漲做了二分類，所以使用 logistic regression，用 50 維 event embedding 的 training data 和 S&P500 漲跌去訓練模型，但或許 50 維中並不是所有參數都有用，因此我們再用 stepwise 雙向學習來做模型的調整，有了這個模型後，再拿模型去預測漲(1)跌(0)並把他們轉成 probability 來供後續的視覺化處理。從分析結果中利用效能指標可以知道模型預測時的準確率達 65.4%，這代表著給我一個今天的新聞標題，我可以藉由其主動受詞的結果來預測明天 S&P500 會不會漲的準確性達 65%。最後 event 的 probability 數值藉由數學運算的轉換變成符合文字雲中的資料型態，利用文字雲可以很直觀的知道，只要在新聞標題中出現某組 event，就容易影響到隔天的 S&P500 上漲的機率。

五、結論

現今對於文字影響股價的研究有許多種，且雖然都用到 NLP 技術但其實原理大相徑庭。例如，有如我們的研究一般，使用事件驅動技術來預測漲跌的研究，也有利用語意框架來判斷文字語意與漲跌關係的研究。在著手做這個專案前花了一些時間讀了不同派別的研究方法，可以說是受益良多，但結果出來後只能說這個模型的可解釋性太差，雖然準確率的表現出乎我意料之外的好（一般相關研究也頂多做到 70%的準確率），但是對於後續分析還是有些力不從心。或許這也是一個我需要努力的方向。

六、參考文獻

Collobert, R., & Weston, J. (2008, July). A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning* (pp. 160-167).Ding, X., Zhang, Y., Liu, T., & Duan, J. (2015, June). Deep learning for event-driven stock prediction. In *Twenty-fourth international joint conference on artificial intelligence*.Socher, R., Chen, D., Manning, C. D., & Ng, A. (2013). Reasoning with neural tensor networks for knowledge base completion. In *Advances in neural information processing systems* (pp. 926-934).Xie, B., Passonneau, R., Wu, L., & Creamer, G. G. (2013, August). Semantic frames to predict stock price ovement. In *Proceedings of the 51st annual meeting of the association for computational linguistics* (pp.873-883).