How Much News Should We Extract For Stock Price Prediction?

Advisor : Jheng-Long Wu
Advisee : Chun-Wei Chang

1. Abstract

翻閱許多文獻及做法之後，我們觀察到針對股價預測的領域鮮少有文獻或研究指出對於文章應取用的數量或是相關的文章篩選方法，且通常對於文章取用的方法多是於字數上的限制，例如每篇文章以前 200 字做為取用目標。本研究提出一個以關聯式方法提取文章，並藉由距離算法得出與目標股票較具關連的文章作為篩選條件。關聯式意味著將所有文章透過 Word Embedding 的方法將其向量化，並利用向量空間上的距離計算方法得到每篇文章與目標股票的相似程度。藉由本次研究，我們的目標在於藉由關聯式方法篩選，為股價預測模型帶來更高的精度。

2. Overview

股市預測一直都是相當熱門的議題，尤其是近年機器學習與深度學習領域的快速發展。其被廣泛用於多個領域，當然股票市場預測是其中一個熱門的應用。當我們在觀測市場時，除了了解過去股價變化、公司財報，還會去關注財經新聞、市場訊息等等，且對於部分的人來說，市場訊息與財經新聞可能是相對重要的一個指標。同理，使用機器學習與深度學習模型預測股價時，常常不僅僅使用歷史股價當作特徵，更會將市場信息如股市新聞當成特徵之一。

至今為止已有有許多研究使用股市新聞做為特徵預測股價，且都獲得了卓越的成效，例如 **Automated news reading: Stock price prediction based on financial news using context-capturing features** *(Michael Hagenau, Michael Liebmann, Dirk Neumann)*，或是利用社群網站預測股市，如: **Twitter mood predicts the stock market.** *(Johan Bollen, Huina Mao, Xiao-Jun Zeng)*。儘管這些論文都有不同的方法去處理文章向量，但是都直接或間接的證明了市場消息之於股價確實是一個有效且強力的特徵。

但是，我們有一個疑問，當預測一個特定上市公司的股價時(e.g.鴻海)，我們如何合理的取得對該公司股價最有影響的市場信息量?將取得的所有市場消息全數使用嗎?全數使用是否會有過多雜訊的問題?還是我們應該取 25%最相關的文章?相反的是不是會有訊息不足的情況?針對這些疑問，本研究設計了一連串的情境，希望經由實驗得知當我們使用市場訊息預測股價時，是否有相對較佳的文章提取數量及方法。

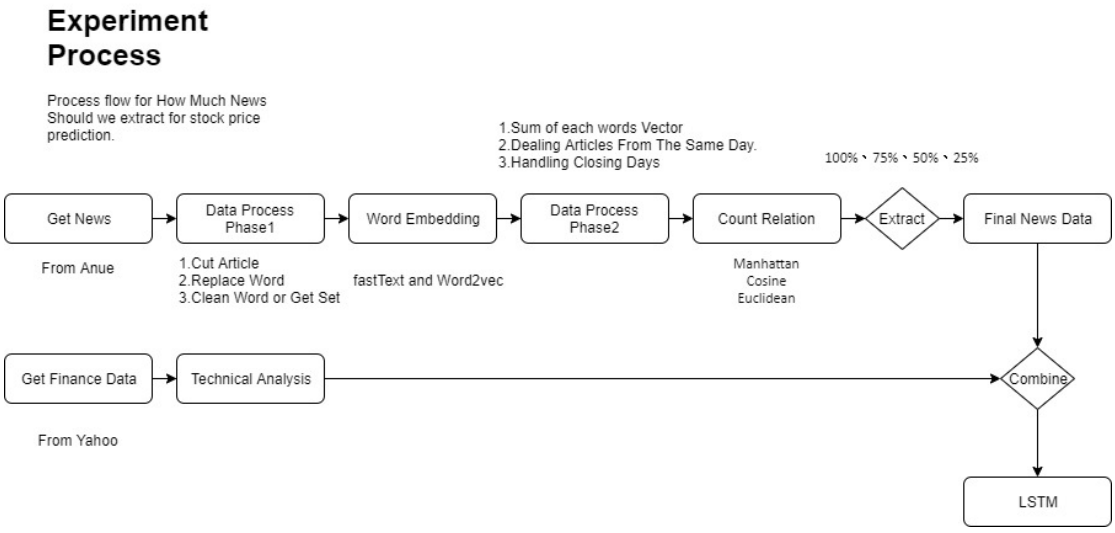
3. Literature Review

對於以文章信息作為預測股價特徵的相關研究有許多，例如於 **Automated news reading: Stock price prediction based on financial news using context-capturing features** *(Michael Hagenau, Michael Liebmann, Dirk Neumann)*此篇文獻中，作者提到了大部分的文本表示方法仍然太倚重 frequency-based 的方法，例如 TF-IDF，且其預測市場的準確率通常沒有超過 60%。可知道現行許多機制都是以詞頻率為主的 TF-IDF 為主。且因傳統 TF-IDF 成效不甚好之緣故，故該篇論文也提出了一個捕捉上下文間的信息之特徵選擇方法，再使用不同的 Feature Representation 的方法實驗。本研究同樣希望聚焦於此，專注於研究是否有較有效的文字信息提取方法，以此作為特徵而精進模型之準確度。

除了使用 TF-IDF 外，還有許多使用分析方法做為新聞處理方法，例如使用情緒分析套件，如同此篇來自 Medium 的教學，或是此專案 Twitter-moods-as-stock-price-predictors-on-Nasdaq 皆是使用情緒指標當作股價預測的特徵，但是此類型的研究並無聚焦於應該取多少文章做為特徵，而通常是全數的文章皆使用。無論是欲使用 TF-IDF 或是情緒分析做為 Feature Representation，同樣都會面臨此次研究的疑問:究竟應該取多少文章來使用會帶來更好的結果?

4. Method

4-1 Process Flow



於上圖中，本研究展示了實驗流程設計。針對新聞資訊，首先從鉅亨網取得財經新聞資料，再來做資料清整，經過 word embedding model 之後，進行第二階段資料清整，清整後即計算關聯程度，並根據不同的門檻值提取關聯文章，且與財經資訊，如股價、技術分析等結合再一起，最後推送至 LSTM Model。下方各小節會做進階說明。

4-2 Relation Method

在過去的實驗中，我們使用關聯式方法(Relation Method)去提取與目標公司股價具有高度關聯的股市信息，並且我們獲得了很好的成效。簡單來說，關聯式方法即是使用 word2vec 或 fastText 的方式將每一個詞轉換為一個向量，並將每一篇文章的詞向量加總以代表一篇文章的向量，而後找到代表目標公司的信息做為中心，再利用距離算法得出每篇文章與中心文章在向量空間的距離，以此表達關聯程度。關於先前的研究，請參考此處 Github。

4-3 Distance Between Each News and Target

承 4-2 所述，在使用 Word2vec 與 fastText 將每一篇新聞及中心文章化為向量之後，我們使用了三種方法來計算每篇文章與中心文章的距離，分別是:

1. 歐式距離(Euclidean Distance)

$$d(x,y) := \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2}$$

2. 曼哈頓距離(Manhattan Distance)

$$d_{12} = \sum_{k=1}^{n} |x_{1k} - x_{2k}|$$

3. 餘弦相似性(Cosine)

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}$$

4-4 Center – Company information

原先本次研究使用公司名稱做為所謂的中心文章，例如鴻海(2317TW)，即使用「鴻海」一詞做為中心文章代表鴻海這支股票。

考量到「鴻海」一詞是否能有效的代表鴻海這間公司，因此本研究還設計了另一種情境:以維基百科之該公司的全部介紹做為中心文章代表該公司的中心文章向量。

4-5 Data Processing

為避免多餘的詞彙造成的雜訊導致距離計算出現誤差，因此本研究使用了停用字辭典，將例如「那麼」、「曾經」等等較不具意義的詞彙於斷詞時去除。此次研究除了停用字典之外，額外考量到即使去除了停用字，會不會仍然有不斷重複的詞造成雜訊過多的問題，因此也設計了取集合的方式將重複的詞去除，但是需特別說明的是，使用集合取不重複值時，會有文字順序打亂的情況，如此的情況對於預測模型是否會有更好的結果，本次研究也相當的有興趣。

另外，在文字資料前處理的部分，本研究事先將所有公司名稱統一，例如將 2317TW、2317、鴻海精密等詞統一變更為鴻海，以此獲得統一的公司名稱。

對於股價、開盤、收盤等等的資料，本次研究使用 Yahoo 股市的歷史股價並且以收盤價(Close)作為研究目標。除了股價、成交量、開盤收盤等等基礎資訊之外，還額外使用了套件計算出 58 種技術指標，包括 ：Accumulation/Distribution Index (ADI)、On-Balance Volume (OBV)、Moving Average Convergence Divergence (MACD)等等，請查看使用之套件。

4-6 Model

因為股價市場乃一具時序性的資料，因此本研究使用了於時序性問題中有效且著名的模型 ：LSTM(Long Short-Term Memory)做為預測模型。

4-7 Architecture

- ✓ 使用模型 ：LSTM
- ✓ Model Setting : Learning Rate : 0.01 ; hidden size = 128 ; timesteps =7 ; Epoch:5000
- ✓ 目標股票 ： 中華電信、仁寶、兆豐金、鴻海、遊戲橘子
- ✓ 資料期間 ：2018/06/14 ~ 2019/07/26
- ✓ 文章資料:來自鉅亨網共 34717 篇
- ✓ 資料切割 ： 訓練集:70%、驗證集:15%、測試集:15%
- ✓ 關聯式提取文章資料數: TOP 100%(全取)、75%、50%、25%
- ✓ 文章文字資料處理方法 ： 停用字、停用字+取 SET
- ✓ Word Embedding Model : Word2vec、fastText
- ✓ 計算距離方法 ： 歐式距離、曼哈頓距離、餘弦相似度

5. Result

5-1 Data Analysis

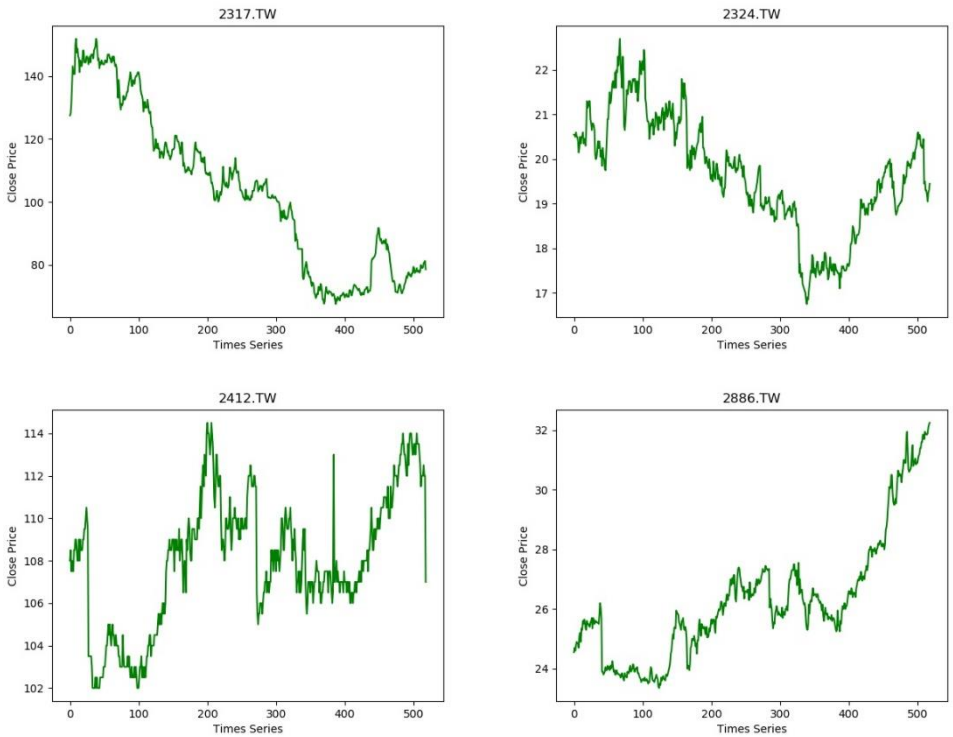
在展示結果之前，必須要先對所有股票的波動及其大約價格有所基本的了

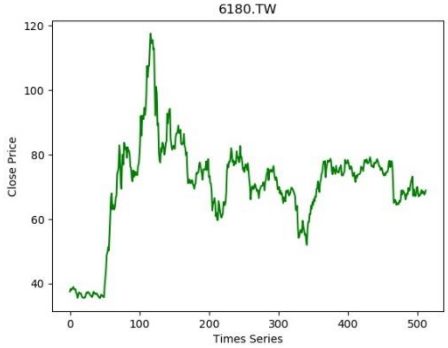
解，因此首先展示統計資訊如表:

Statistics – Stock Price(Close)					
	遊戲橘子 (6180TW)	鴻海 (2317TW)	仁寶 (2324TW)	中華電信 (2412TW)	兆豐金 (2886TW)
Mean	70.4	103.6	19.7	108	26.3
Max	117.5	151.9	22.7	114.5	32.3
Min	35.5	67.6	16.8	102	23.4
Max-Min	82.5	84.3	5.9	12.5	8.9
Standard deviation	14.73	25.34	1.3	3.1	2.1
Class1-Stock Price	股價中間	股價偏高	股價偏低	股價偏高	股價偏低
Class2-fluctuation	波動較大	波動較大	波動較小	波動較小	波動較小

(表 1，股價資訊基本統計)

由表 1 可以清楚地發現，遊戲橘子及鴻海於資料期間有巨大的波動，而仁寶、中華電信及兆豐金則可歸類於相對穩定的股票。其中鴻海、中華電信屬於股價相對偏高的；遊戲橘子則居中，而仁寶及兆豐金則是相對較低。





(圖 1-圖 5，各股票收盤價之變化)

於圖 1 至圖 5，本研究展示了各個股票於此資料期間的收盤價波動，從 y 軸 Close Price 之區間即可看出各個股票於資料期間內波動的多寡。

5-2 Experiment Design Introduction

如同 4-3 章節所介紹，本研究主體上分為兩個方法決定所謂的中心文章，一為以公司名稱作為中心文章(以下簡稱中心)；二為以維基百科之公司介紹做為中心，在此小節中將詳細說明實驗流程。

首先，將爬取下來的新聞資料，經過斷詞之後以 Word2vec 及 fastText 分別訓練所有文章，藉此可得到每一個詞所代表的向量。然而，考量到雜訊的問題，因此事先將文章做兩種處理:其一為僅使用停用辭典(即那麼、好像等詞)清理文章，其二為除了使用停用辭典外，再將文章以取集合的方式去除重複詞。處理完後將一篇文章中剩餘的詞轉為向量，並將之全部相加，以此代表一篇文章之向量。

獲取每篇文章的向量之後，將要決定每篇文章與中心的關聯程度，因此分別以歐式距離、曼哈頓距離、餘弦相似度公式計算得出關聯程度。接著便是本次實驗的重點:應該取多少數量的文章做為特徵，能為模型創造更好的結果?因此本次研究設計了四個門檻，分別是關聯度最高之前 25%、50%、75%以及全部文章皆使用四種情境。

此時還面對一個問題，即使是文章數量最精簡的取前 25%最高的關聯程度，也仍然存在一天有多篇文章的情況，本次研究設計將當天所有取出來的文章相加做為該日新聞文章特徵以此預測股價。另外，還須解決股市未開盤的情形，本次實驗將未開盤當日之資訊，加到前一個股市開盤日，舉例來說，星期六日皆未開盤，但當日也仍會有財經相關文章，本研究將之加到星期五的文章向量之中，藉此來預測下一個開盤日，也就是星期一的股價。

5-3 Scenario1-Company name as center

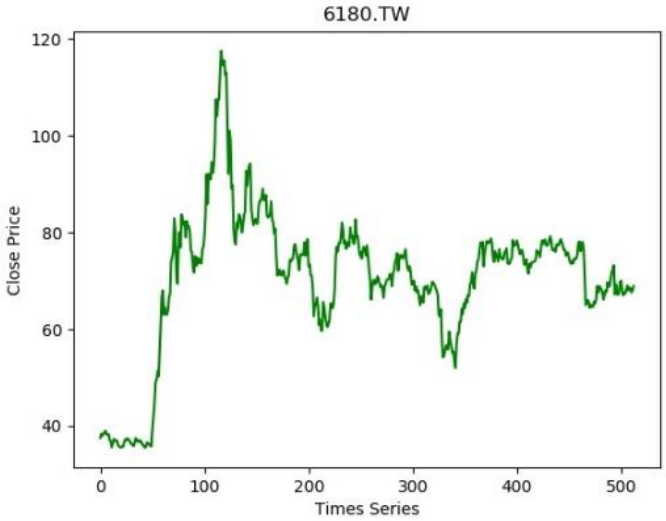
於此小節，將展示以公司名稱做為中心的結果。

遊戲橘子(6180TW)				
	停用字+取集合 Word2vec	**停用字+取集合 fastText**	**停用字 Word2vec**	**停用字 fastText**
cos100	5.769372463	4.20242691	4.427407742	**3.672474146**
cos75	5.160563469	**3.832211256**	4.376977921	5.183384895
cos50	3.852986813	4.207755566	5.623907566	5.500298977
cos25	4.602060795	6.493478298	4.148766041	4.488929272
euc100	3.894173145	4.529332638	4.344653606	5.439959526
euc75	6.360085011	4.917878628	4.46331358	5.78036499
euc50	5.276314735	3.939327478	5.207735538	4.368473053
euc25	**3.701768875**	4.077071667	4.073010921	4.56970644
manha100	3.71958518	4.118139744	5.69099474	4.364296436
manha75	5.68791151	7.105129242	4.459958553	4.439599037
manha50	4.765241623	4.99917078	4.750805378	5.216396809
manha25	4.980922699	5.586363792	**3.901718855**	5.211099148
Mean	4.81424886	4.834023833	4.622437537	4.852915227

(表 2-遊戲橘子結果)

請查看表 2，於開頭介紹一下表格資訊，首先上方 Column 即為停用字與取集合是否，及分別使用 Word2vec 與 fastText 的結果；左方則代表關聯度計算方法及取用最具關連的文章之百分比，cos 代表餘弦相似度；euc 代表歐式距離；manha 則代表曼哈頓距離。例如 cos100 則代表使用餘弦相似度計算關聯程度，並全部文章都使用未做篩選；euc25 則代表使用歐式距離計算關聯程度，且僅取前 25%關聯程度最高的文章。紅字代表該方法中結果最好的做法，而所有數值皆是預測測試集後與真實股價的絕對值取平均而來。

由表 2 可以見得各個情境中可以見得分別由 euc25、cos75、manha25 及 cos100 結果最好，其中又以僅有停用字與 fastText 的組合且使用餘弦相似度的文章全取的方法結果最好。



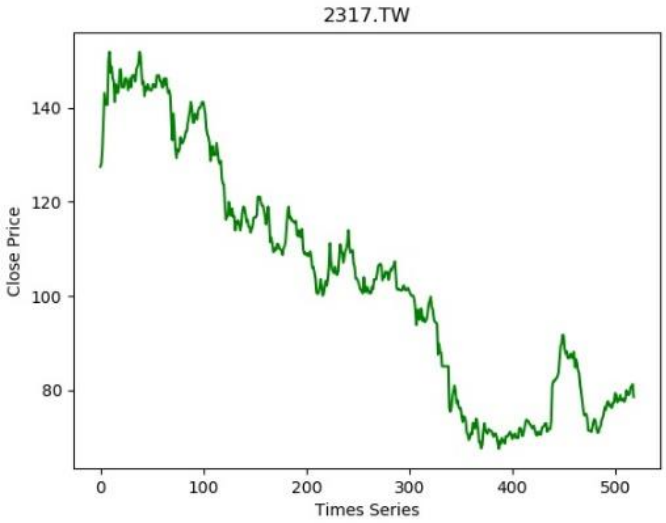
(圖 6 -遊戲橘子股價)

從由遊戲橘子的股價，可得知橘子在訓練資料期間是屬於波動較大的，整體股價從最低 30 多一路到逼近 120，遊戲橘子的股價於此次實驗的五檔來說股價算是中間等級。橘子在測試資料的期間(圖 6 末端) 比較平穩，因此預測股價的結果以四個方法中最好的平均誤差可在 3-4 之間。接著需觀察其他實驗股票的結果，檢驗是否能到最適合各個股票特性的文章提取方法。

鴻海(2317TW)				
	停用字+取集合 Word2vec	**停用字+取集合 fastText**	**停用字 Word2vec**	**停用字 fastText**
cos100	10.0695343	11.01199627	12.18185711	8.946799278
cos75	10.95836926	12.07090664	11.44894028	11.2600317
cos50	9.178712845	10.10229397	**7.974783897**	11.07693481
cos25	9.423544884	13.637784	10.05056286	**7.297438622**
euc100	9.203995705	11.85216713	9.941425323	8.884176254
euc75	10.65297604	**6.491306305**	8.702059746	7.982998848
euc50	10.42811108	7.638343811	10.1912508	10.76142788
euc25	9.792844772	10.48919201	11.23455906	12.37504005
manha100	10.02131367	9.507936478	8.864254951	10.59237576
manha75	**7.771401882**	9.745084763	12.52011967	10.72148609
manha50	9.092220306	8.421927452	9.848248482	9.450713158
manha25	11.76803207	10.61561203	9.773359299	9.816208839
Mean	9.8634214	10.1320459	10.22761846	9.930469275

(表 3-鴻海結果)

表 3 中可見得鴻海最好的結果分別是 manha75、euc75、cos50、cos25，其中又以取集合加上停用字且使用 fastText 之 euc75 結果最好。從鴻海的結果當中可以發現最好四個文章提取模式及數量中已無文章全部取用的方法。



(圖 7 -鴻海股價)

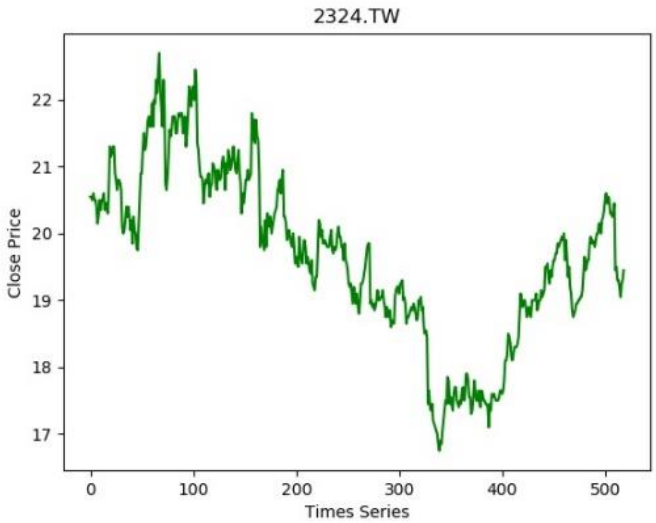
鴻海的類別歸類在股價偏高，在波動上與遊戲橘子同樣為波動較大，但遊戲橘子的股價是不斷上下波動的，而鴻海在訓練集上屬於不斷向下的趨勢。在這樣的情況下，集合、停用字與 fastText 的 euc75 效果最好。本研究將會在此小節末尾統計各股價的特性及其結果。

仁寶(2324TW)				
	停用字+取集合 Word2vec	**停用字+取集合 fastText**	**停用字 Word2vec**	**停用字 fastText**
cos100	2.549031258	7.918372154	6.110300064	3.430308104
cos75	**2.144803286**	2.176797628	4.016634941	3.781284571
cos50	6.671281338	3.962017059	4.18351984	2.69819212
cos25	3.601208687	2.678389788	2.453976631	2.758422375
euc100	3.804149389	5.252024174	3.016422272	**1.20079422**
euc75	3.211007833	6.185795784	10.53308868	3.009369612
euc50	4.733278751	2.26322484	3.595303297	4.436717033
euc25	6.089751244	2.565932751	5.880330086	2.66621232
manha100	3.294461966	**1.909045935**	**2.172570467**	3.231724739
manha75	2.791586399	4.883197308	3.75864315	5.030211449
manha50	5.593880177	4.851946831	2.708052874	2.885242462
manha25	2.470824003	8.734406471	3.737259388	3.405486584

Mean	3.912938694	4.448429227	4.347175141	3.211163799

(表 4-仁寶結果)

仁寶各個情境中結果最好的分別是 cos75、manha100、manha100、euc100，其中又以僅有停用字與 fastText 情況下的 euc100 結果最好，預測平均誤差甚至壓在 1.5 以內。不過這次還有一個有趣的結果:對於仁寶來說，似乎文章信息全部取用的狀況下容易會有更好的結果。



(圖 8 -仁寶股價)

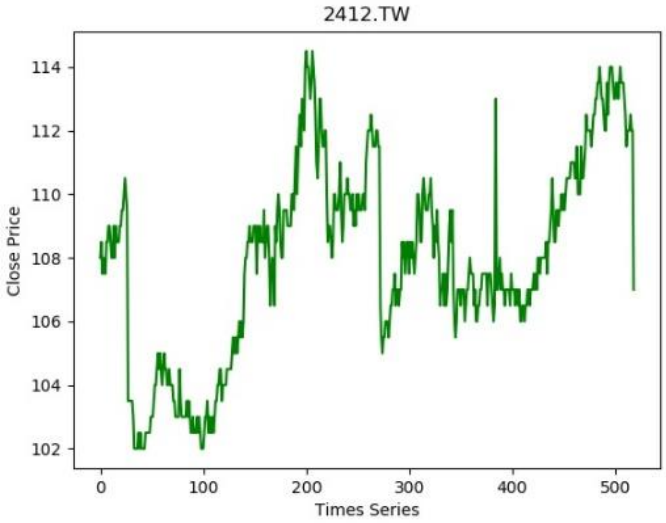
從圖 8 的仁寶股價中，觀察 y 軸可以發現仁寶確實算是相當穩定的股票，從結果上來看，是否可以得到相對穩定的股票所有文章皆須使用，才會用更佳的結果?對於這個推論，還需要其他實驗來確認，因此需繼續觀察其餘股票的結果。

中華電(2412TW)				
	停用字+取集合 Word2vec	**停用字+取集合 fastText**	**停用字 Word2vec**	**停用字 fastText**
cos100	11.00144863	6.303227425	8.81043148	9.160165787
cos75	7.405096054	8.811629295	10.36863995	9.622509003
cos50	6.39262867	**5.963294029**	10.86762524	7.833959103
cos25	**6.044153214**	13.41850662	12.18742275	11.64003658
euc100	8.239095688	10.52468204	9.277729034	10.0938015
euc75	9.887597084	6.10874939	11.16864204	**7.236570835**
euc50	6.994842052	7.406172276	**6.397054195**	13.0818882

euc25	10.65661812	9.342600822	7.570558071	8.516758919
manha100	8.247337341	10.63753891	8.064955711	7.903925896
manha75	9.213358879	7.609258652	11.54217529	9.636475563
manha50	9.488128662	8.410070419	9.340592384	14.74173927
manha25	8.962533951	9.939730644	11.4704113	8.617750168
Mean	8.544403195	8.706288377	9.755519787	9.840465069

(表 5-中華電結果)

中華電信結果最好的分別是 cos25、cos50、euc50、euc75，其中以同時使用停用字與取集合，且使用 fastText 的方法，cos50 的情況下最好。



(圖 9-中華電信股價)

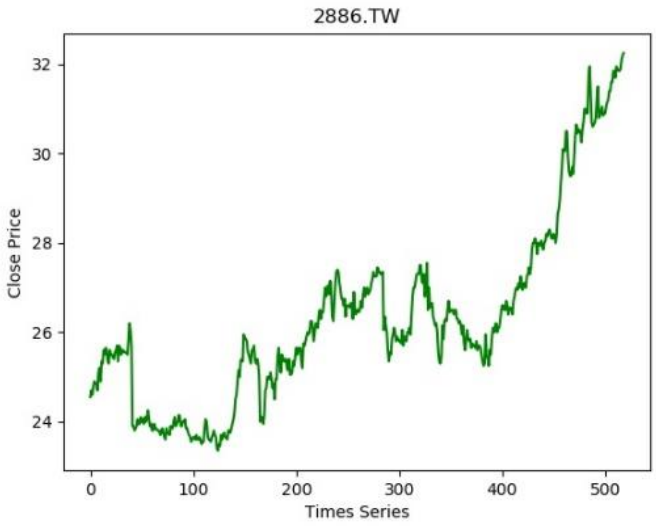
中華電信的股價在分類上屬於波動較小；股價偏高。中華電信的預測結果無全部文章取用的情況，以此情境來說，波動較小股價偏高的情境使用較為精簡的信息會帶來更好的成果。

兆豐金(2886TW)				
	停用字+取集合 Word2vec	**停用字+取集合 fastText**	**停用字 Word2vec**	**停用字 fastText**
cos100	**4.972481251**	10.45316696	12.32521629	10.1850214
cos75	10.88432598	9.944187164	11.31442928	9.833803177
cos50	9.313117027	**5.564670563**	11.25989914	9.73268795
cos25	17.77569962	8.126306534	12.89756012	9.501477242
euc100	11.2406168	9.207251549	10.73659992	8.281515121
euc75	7.543545246	8.601363182	9.497119904	**5.526361465**

euc50	14.17605591	12.35475731	10.45229435	8.563128471
euc25	9.899985313	9.119620323	14.10876656	10.3940773
manha100	9.371765137	9.286621094	13.38238144	11.69591236
manha75	13.44291592	16.21824265	**7.020232201**	10.49667358
manha50	10.99926567	10.86186028	13.91496181	13.40153027
manha25	11.90304661	12.61801147	11.67505455	10.13368988
Mean	10.96023504	10.19633826	11.54870963	9.812156518

(表 6-兆豐金結果)

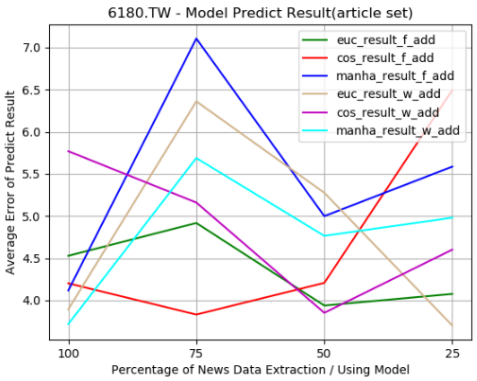
兆豐金結果最好的分別是 cos100、cos50、manha75、euc75，其中以同時使用停用字與取集合，且使用 word2vec 的方法，cos100 的情況下最好。

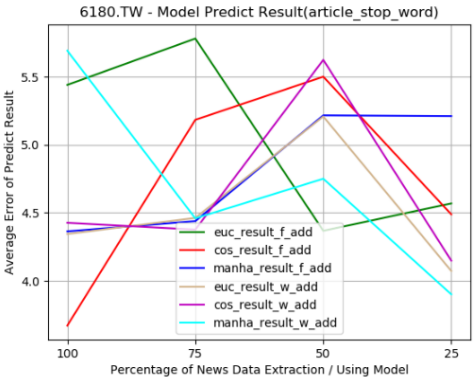


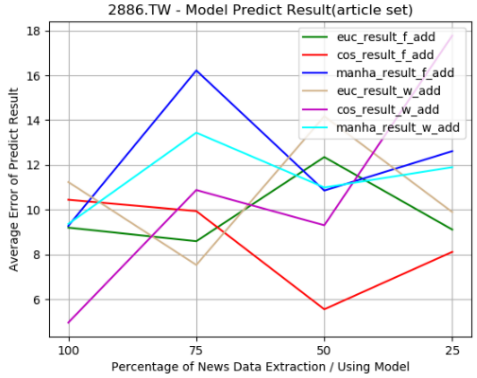
(圖 10 -兆豐金股價)

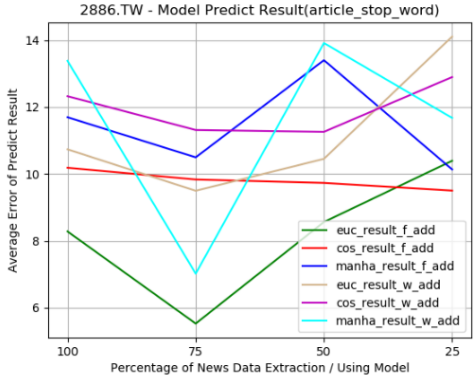
兆豐金的股價特性相較來說屬於波動較小，股價較低的，且從資料期間可見得兆豐金的股價呈現向上的趨勢。在這樣的情境下，最好的文章提取方法結果出現在 cos100 的情況。

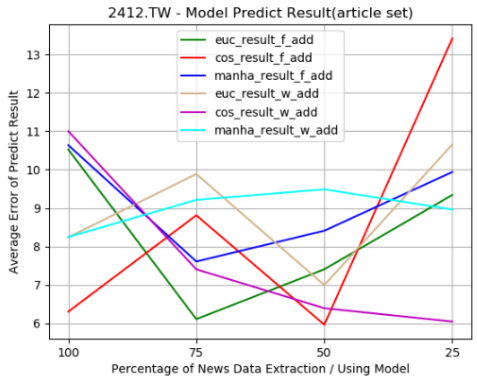
此小節一一介紹完各個股價及其情境的結果，但至此資訊量過大因此較不便進行分析，因此本研究於下方將會統計相關資訊。

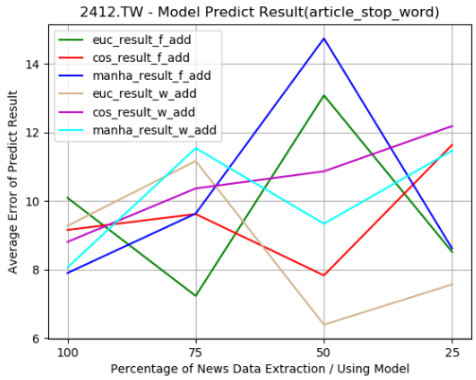

6180.TW - Model Predict Result(article set)
euc_result_f_add
cos_result_f_add
manha_result_f_add
euc_result_w_add
cos_result_w_add
manha_result_w_add
Average Error of Predict Result
Percentage of News Data Extraction / Using Model

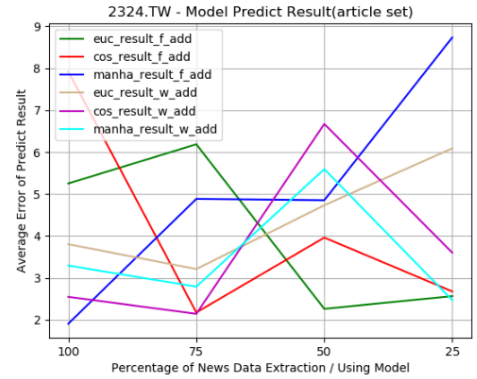

6180.TW - Model Predict Result(article_stop_word)
euc_result_f_add
cos_result_f_add
manha_result_f_add
euc_result_w_add
cos_result_w_add
manha_result_w_add
Average Error of Predict Result
Percentage of News Data Extraction / Using Model

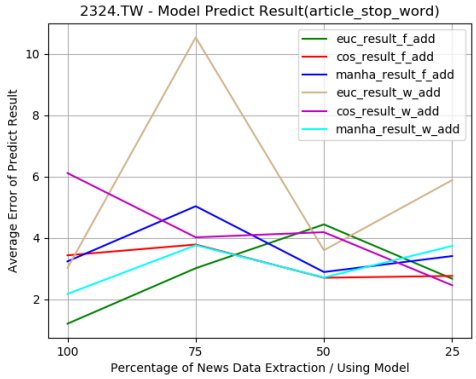

2886.TW - Model Predict Result(article set)
euc_result_f_add
cos_result_f_add
manha_result_f_add
euc_result_w_add
cos_result_w_add
manha_result_w_add
Average Error of Predict Result
Percentage of News Data Extraction / Using Model

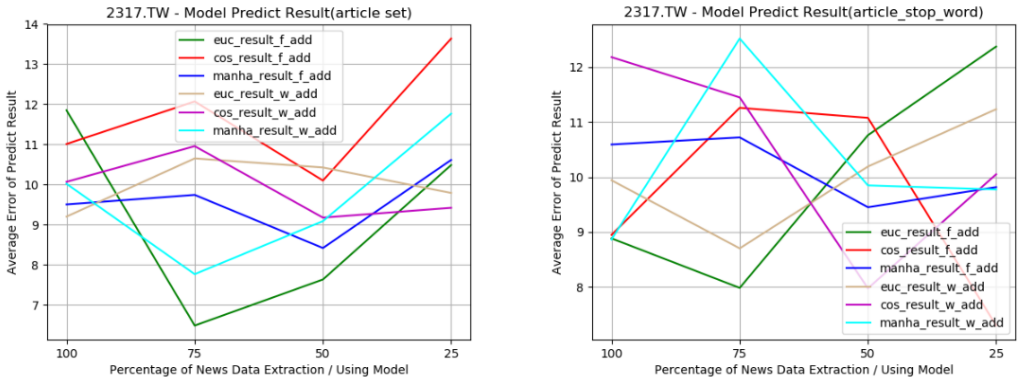

2886.TW - Model Predict Result(article_stop_word)
euc_result_f_add
cos_result_f_add
manha_result_f_add
euc_result_w_add
cos_result_w_add
manha_result_w_add
Average Error of Predict Result
Percentage of News Data Extraction / Using Model


2412.TW - Model Predict Result(article set)
euc_result_f_add
cos_result_f_add
manha_result_f_add
euc_result_w_add
cos_result_w_add
manha_result_w_add
Average Error of Predict Result
Percentage of News Data Extraction / Using Model


2412.TW - Model Predict Result(article_stop_word)
euc_result_f_add
cos_result_f_add
manha_result_f_add
euc_result_w_add
cos_result_w_add
manha_result_w_add
Average Error of Predict Result
Percentage of News Data Extraction / Using Model


2324.TW - Model Predict Result(article set)
euc_result_f_add
cos_result_f_add
manha_result_f_add
euc_result_w_add
cos_result_w_add
manha_result_w_add
Average Error of Predict Result
Percentage of News Data Extraction / Using Model


2324.TW - Model Predict Result(article_stop_word)
euc_result_f_add
cos_result_f_add
manha_result_f_add
euc_result_w_add
cos_result_w_add
manha_result_w_add
Average Error of Predict Result
Percentage of News Data Extraction / Using Model



(圖 11-20-各股價之各情境結果變化圖)

圖 11-20 將此小節中的表格做視覺化，圖片的 title 為股價以及其使用停用字與取集合(article set)或是僅使用停用字(article_stop_word)。標示代表計算距離的方法，f 則代表使用 fastText 的情境，反之 w 代表使用 word2vec 的情境，而最後方的 add 則是表示向量處理方法，意指將一篇文章中的所有詞彙向量加總代表一篇文章，並且將每日多篇文章也做加總。X 軸則代表 4 個分類，分別是全部文章皆取用(100%)、取用前 75%、50%及 25%最具關聯的文章，y 軸則代表了於測試資料期間的預測值與真實值平均誤差。

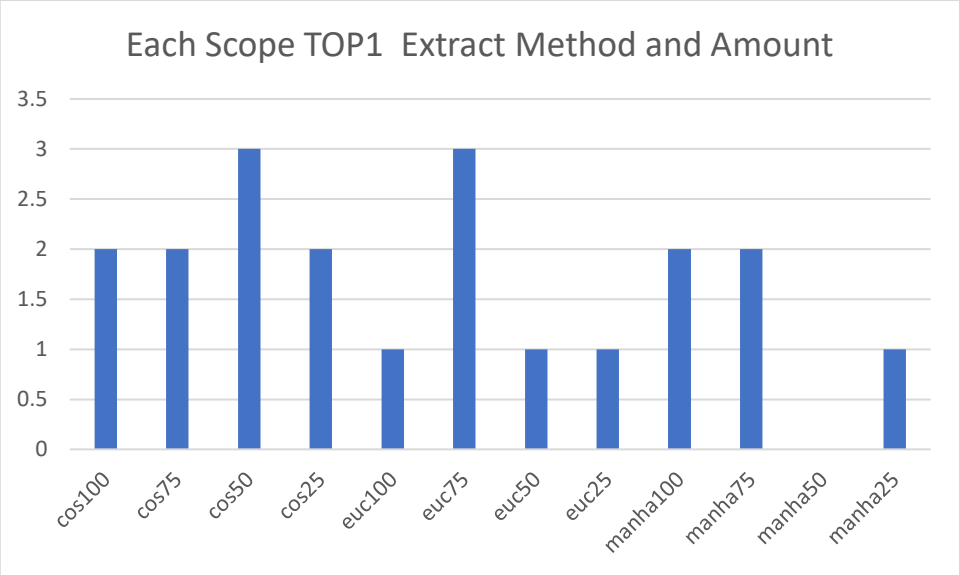
從結果的變化上來看，並未發現明顯且有效的規律，設計此實驗時本是希望驗證以關聯式方法提取新聞文章，並藉由此方法取得與目標公司股價較有關係之新聞作為特徵之一。

Result Table					
	遊戲橘子 (6180TW)	**鴻海 (2317TW)**	**仁寶 (2324TW)**	**中華電信 (2412TW)**	**兆豐金 (2886TW)**
Class1-Stock Price	股價中間	股價偏高	股價偏低	股價偏高	股價偏低
Class2-fluctuation	波動較大	波動較大	波動較小	波動較小	波動較小
Best Result-Method	Stopword (fastText)	Stopword+Set (fastText)	Stopword (fastText)	Stopword+Set (fastText)	Stopword+Set (word2vec)
Best Result-News Extract	cos100	euc75	euc100	cos50	cos100

(表 7-結果比較)

從表 7 的結果整理中。以 Best Result Method 來說，並未發現將文章詞彙多做一步取集合的動作會帶來更好的成效，以目前的結果來看，兩者各在不同的

情境下各有優劣。唯一稍微有點線索的是使用 fastText 作為 Word Embedding 的方法較容易在這個情境假設下帶來好的結果。



(圖 21 -各股票各個情境中前三最優秀的新聞提取方法)

根據上面的所有統計，本研究整理了各個情境中表現最好的模式(各表中的紅字)，可知在各個情況有其不一樣的新聞提取方法，並沒有一定的答案說明餘弦、曼哈頓或歐基里德孰者一定是最好；亦或是何種類型的股票應當用何種的文章處理規則。

但是僅看到這裡，仍然無法下定論。或許是單純以公司名稱當作中心點的情況不足以代表某家公司的資訊，因此仍須看第二種情境:以該公司維基百科的介紹作為中心點。

5-4 Scenario2-Company introduction(wiki) as center

於此小節中，將展示以維基百科所獲取之公司介紹作為中心文章的結果。在此僅先快速帶過每一個實驗的結果，於末尾會做統計，下一小節將會比較兩者之差別

遊戲橘子(6180TW)				
	停用字+取集合 Word2vec	**停用字+取集合 fastText**	**停用字 Word2vec**	**停用字 fastText**
cos100	5.471982956	4.607812405	4.342301846	**3.743516922**
cos75	5.304370403	4.171423912	4.053124905	4.590974808
cos50	3.633210421	4.705266476	5.520480633	5.606752872
cos25	4.687556744	5.95623064	4.085755825	4.218563557
euc100	3.757919312	4.632584095	4.088177204	6.324942112

euc75	7.038825512	5.125324249	4.293293953	5.808125496
euc50	4.339300632	**3.555383682**	5.595999241	4.254151821
euc25	4.312766552	4.092006683	4.167444229	5.556242466
manha100	**3.607044935**	4.407214642	5.871804237	4.413272858
manha75	5.687188148	7.015464306	**3.967972994**	4.660482883
manha50	4.921051025	5.131742001	4.99993515	5.217596054
manha25	4.741965771	5.301196098	3.987798214	5.491611481
Mean	4.791931868	4.891804099	4.581174036	4.990519444

(表 8-遊戲橘子結果)

以公司維基百科介紹做為中心的實驗中，遊戲橘子結果最好的分別是 manha100、euc50、manha75、cos100，其中以同時使用停用字與取集合，且使用 fastText 的方法，euc50 的情況下最好。

如同上一小節所述，遊戲橘子乃是屬於股價中間，波動較大的類型，在這樣的情況下，歐幾里得距離及取 top50%的文章將會帶來更好的成效。觀察其表現最好的四個部分，其提取之新聞量貌似都偏高。

鴻海(2317TW)				
	停用字+取集合 Word2vec	**停用字+取集合 fastText**	**停用字 Word2vec**	**停用字 fastText**
cos100	9.753037453	10.96025944	12.18187332	8.954769135
cos75	10.9023695	11.93570042	11.33180809	11.29585838
cos50	9.226429939	10.14083862	**7.932420731**	11.68110657
cos25	9.540410995	13.32037926	9.728330612	**7.38868475**
euc100	10.14298058	11.77615166	9.88121891	8.952735901
euc75	10.92900753	**6.474448681**	8.571341515	7.893271923
euc50	10.467062	7.658563137	10.15054798	10.70267868
euc25	9.121781349	10.84395885	11.2288332	12.13157463
manha100	10.05418491	9.655977249	8.799819946	10.62920952
manha75	**7.738941669**	9.817399025	11.29621887	10.6709156
manha50	9.514590263	8.260075569	10.08976364	9.356116295
manha25	12.16314697	10.67835903	9.870954514	9.81155014
Mean	9.962828596	10.12684258	10.08859428	9.955705961

(表 9-鴻海結果)

鴻海結果最好的分別是 manha75、euc75、cos50、cos25，其中以同時使用停用字與取集合，且使用 fastText 的方法，euc75 的情況下最好。

鴻海在分類上屬於股價偏高、波動較大的類型，在此情境中以歐基里德距

離取前 75%最具關聯度的文章效果最好。而其表現最好之提取的新聞量分布從 25% 至 75% 皆有。

仁寶(2324TW)				
	停用字+取集合 Word2vec	停用字+取集合 fastText	停用字 Word2vec	停用字 fastText
cos100	2.611529589	7.599591255	5.916483402	3.457205296
cos75	**2.254271984**	**2.303181171**	3.991357088	3.761039019
cos50	6.741624832	3.940158606	4.224511147	2.85285759
cos25	3.558472157	2.836359501	2.572300196	2.679214716
euc100	4.774769306	5.315577984	3.014137268	**1.346060157**
euc75	3.257733822	6.670510769	10.58532524	2.541641951
euc50	5.143311501	2.265646696	3.717725039	4.415756702
euc25	5.91354084	2.462999582	5.518967152	2.674994946
manha100	3.308227539	2.30801177	**2.238997698**	3.096187592
manha75	2.803112268	4.910300732	3.632892847	4.786781788
manha50	5.718047142	4.892970562	2.63019371	2.616176367
manha25	2.303944826	8.695145607	3.778491974	3.34959507
Mean	4.03238215	4.51670452	4.318448563	3.131459266

(表 10-仁寶結果)

仁寶結果最好的分別是 cos75、cos75、manha100、euc100，其中以僅使用停用字，且使用 fastText 的方法，euc100 的情況下最好。

仁寶在分類上屬於股價偏低、波動較小的類型，在此情境中以歐基里德距離文章全部取用的情境效果最好。仁寶這種股價偏低且波動小的股票，新聞取用的較大量會較有更好的成效。

中華電(2412TW)				
	停用字+取集合 Word2vec	停用字+取集合 fastText	停用字 Word2vec	停用字 fastText
cos100	10.74765491	6.2992239	16.35493279	8.612108231
cos75	7.176634312	8.451155663	10.26514053	9.536005974
cos50	**6.176064968**	**5.834035873**	10.99949646	7.900426388
cos25	6.280389309	13.26721382	11.98111057	11.79667568
euc100	8.196774483	10.66388988	7.172255516	10.49454212
euc75	9.307670593	10.91542053	11.66414642	**7.804506779**
euc50	7.189486504	7.449413776	**6.480576038**	13.86944389

euc25	10.74873924	9.348640442	9.118067741	8.226686478
manha100	8.464487076	11.32399464	8.171452522	8.914365768
manha75	8.994543076	7.587309837	10.51533127	10.00335026
manha50	9.290781021	10.28605556	8.962399483	14.99575901
manha25	9.417494774	9.938735008	11.34272099	8.190680504
Mean	8.499226689	9.280424078	10.25230253	10.02871259

(表 11-中華電結果)

中華電結果最好的分別是 cos50、cos50、euc50、euc75，其中以使用停用字與取集合，且使用 fastText 的方法，cos50 的情況下最好。

中華電在分類上屬於股價偏高、波動較小的類型，在此情境中以餘弦距離文章取用 TOP 50%最具關聯的為好。而中華電的結果都集中在 50%及 75%。

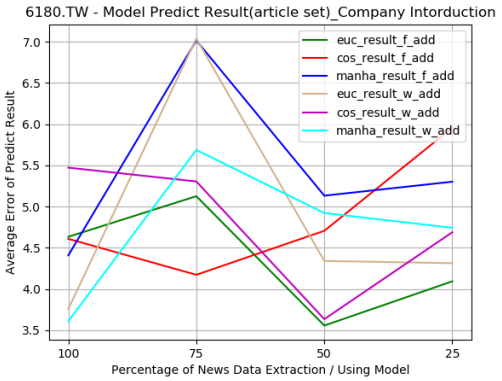
兆豐金(2886TW)				
	停用字+取集合 Word2vec	**停用字+取集合 fastText**	**停用字 Word2vec**	**停用字 fastText**
cos100	**4.920205116**	10.42387772	12.43729687	10.93248653
cos75	11.01423931	10.01235485	11.07069492	8.262019157
cos50	9.291344643	**8.232339859**	11.19854832	10.51749802
cos25	17.90155029	9.091210365	12.92625332	9.409583092
euc100	11.2406168	9.275197029	11.02719212	7.806950569
euc75	7.561427593	8.553201675	9.479948997	**5.289246559**
euc50	13.99544048	12.28599834	9.96972847	8.911844254
euc25	9.835394859	9.105512619	13.85581875	8.850128174
manha100	9.373638153	9.288192749	13.46989918	8.922638893
manha75	13.53618336	16.15042496	**8.528308868**	10.62388515
manha50	11.23235321	10.88878345	15.30748844	13.45588017
manha25	11.90568638	12.68269348	9.894987106	10.48612785
Mean	10.98400668	10.49914893	11.59718045	9.455690702

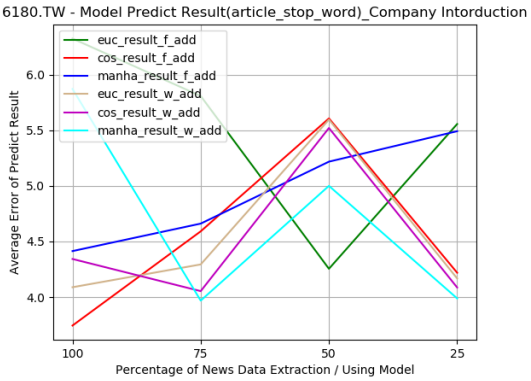
(表 12-兆豐金結果)

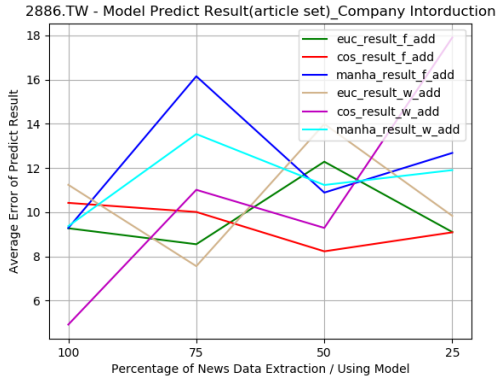
兆豐金結果最好的分別是 cos100、cos50、manha75、euc75，其中以使用停用字與取集合，且使用 word2vec 的方法，cos100 的情況下最好。

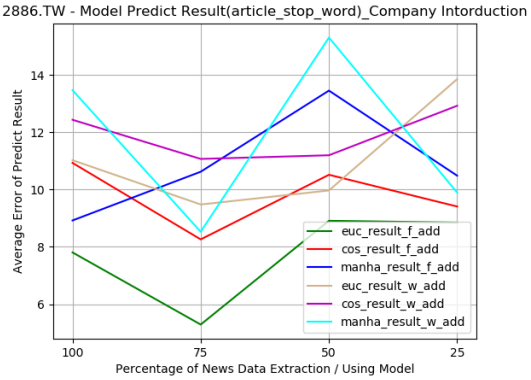
中華電在分類上屬於股價偏低、波動也較小的類型，在此情境中以餘弦距離文章全取為好。

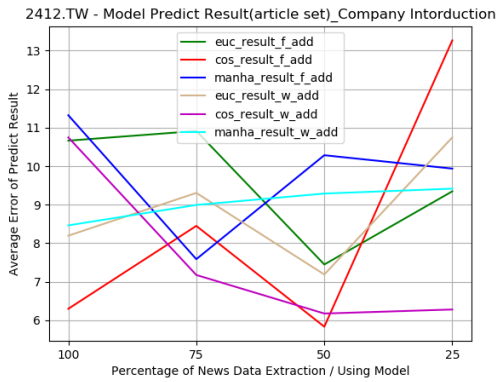
以上的表格僅是快速示意，接著將有視覺化圖表協助理解:

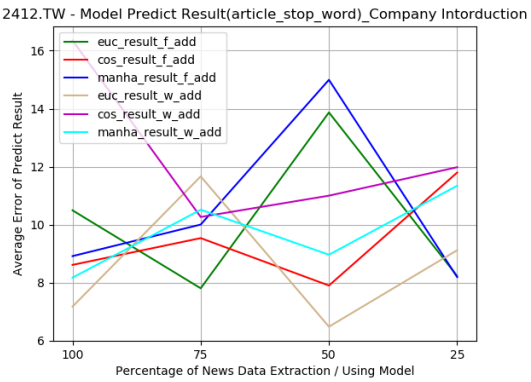

6180.TW - Model Predict Result(article set)_Company Intorduction
euc_result_f_add
cos_result_f_add
manha_result_f_add
euc_result_w_add
cos_result_w_add
manha_result_w_add
Average Error of Predict Result
Percentage of News Data Extraction / Using Model

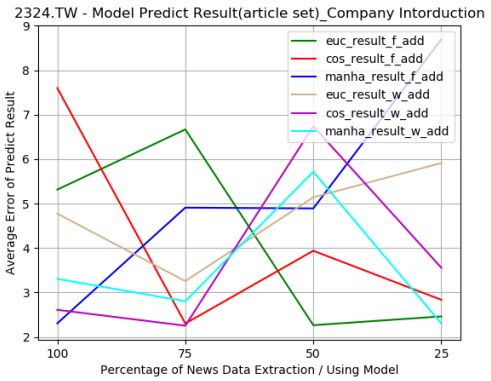

6180.TW - Model Predict Result(article_stop_word)_Company Intorduction
euc_result_f_add
cos_result_f_add
manha_result_f_add
euc_result_w_add
cos_result_w_add
manha_result_w_add
Average Error of Predict Result
Percentage of News Data Extraction / Using Model

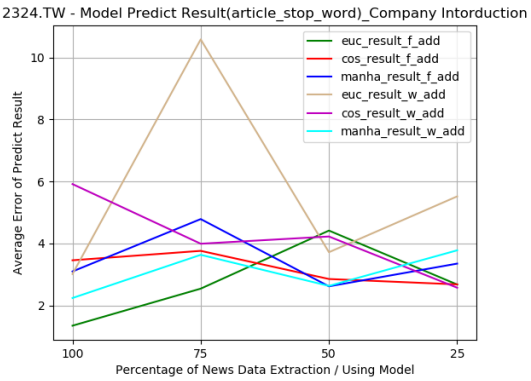

2886.TW - Model Predict Result(article set)_Company Intorduction
euc_result_f_add
cos_result_f_add
manha_result_f_add
euc_result_w_add
cos_result_w_add
manha_result_w_add
Average Error of Predict Result
Percentage of News Data Extraction / Using Model

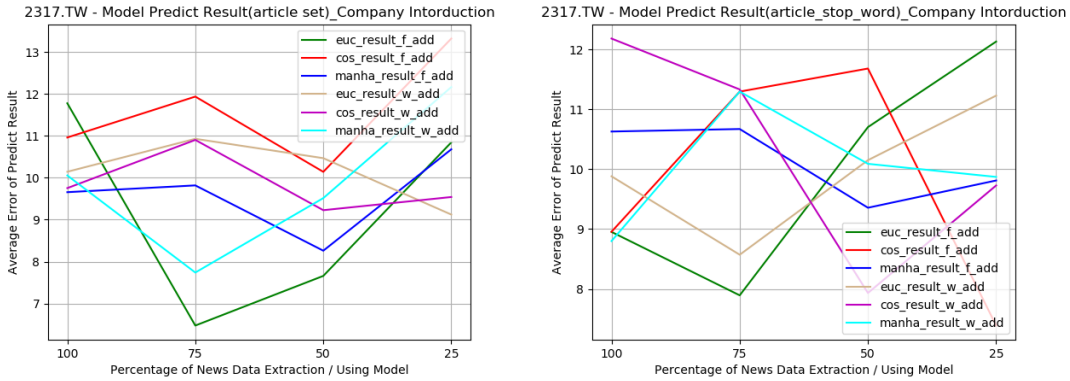

2886.TW - Model Predict Result(article_stop_word)_Company Intorduction
euc_result_f_add
cos_result_f_add
manha_result_f_add
euc_result_w_add
cos_result_w_add
manha_result_w_add
Average Error of Predict Result
Percentage of News Data Extraction / Using Model


2412.TW - Model Predict Result(article set)_Company Intorduction
euc_result_f_add
cos_result_f_add
manha_result_f_add
euc_result_w_add
cos_result_w_add
manha_result_w_add
Average Error of Predict Result
Percentage of News Data Extraction / Using Model


2412.TW - Model Predict Result(article_stop_word)_Company Intorduction
euc_result_f_add
cos_result_f_add
manha_result_f_add
euc_result_w_add
cos_result_w_add
manha_result_w_add
Average Error of Predict Result
Percentage of News Data Extraction / Using Model


2324.TW - Model Predict Result(article set)_Company Intorduction
euc_result_f_add
cos_result_f_add
manha_result_f_add
euc_result_w_add
cos_result_w_add
manha_result_w_add
Average Error of Predict Result
Percentage of News Data Extraction / Using Model


2324.TW - Model Predict Result(article_stop_word)_Company Intorduction
euc_result_f_add
cos_result_f_add
manha_result_f_add
euc_result_w_add
cos_result_w_add
manha_result_w_add
Average Error of Predict Result
Percentage of News Data Extraction / Using Model



(圖 22-31-各股價之各情境結果變化圖)

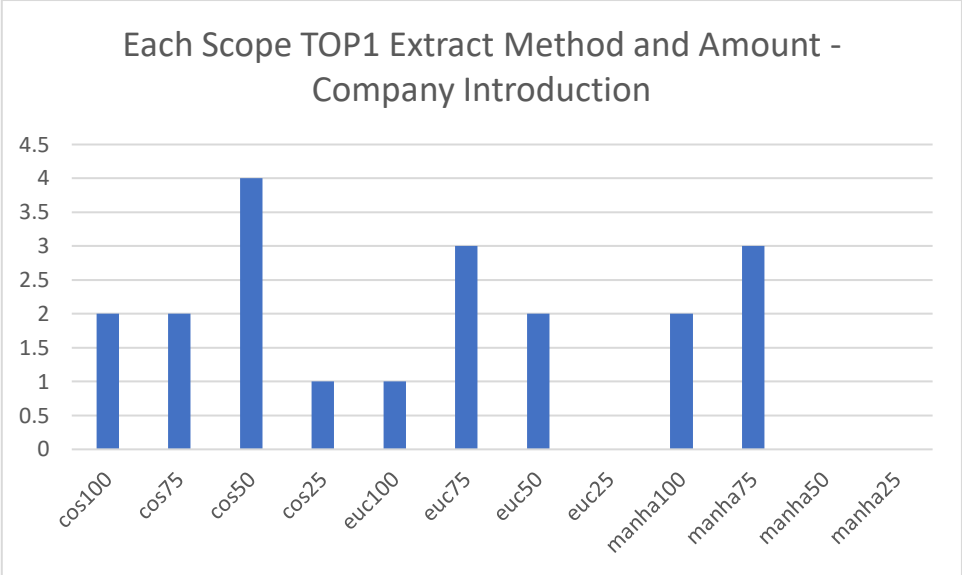
以公司簡介做為中心文章的情境來說，與以公司名稱作為中心文章的比對來看(p12-13)，其實兩者的走勢相當的相似，同樣從結果的變化中並未發現明顯且有效的規律。但是以視覺化的圖表觀察似乎過於籠統，下方最做統整的表格。

Result Table(Company Introduction)					
	遊戲橘子 (6180TW)	**鴻海 (2317TW)**	**仁寶 (2324TW)**	**中華電信 (2412TW)**	**兆豐金 (2886TW)**
Class1-Stock Price	股價中間	股價偏高	股價偏低	股價偏高	股價偏低
Class2-fluctuation	波動較大	波動較大	波動較小	波動較小	波動較小
Best Result-Method	Stopword+Set (fastText)	Stopword+Set (fastText)	Stopword (fastText)	Stopword+Set (fastText)	Stopword+Set (word2vec)
Best Result-News Extract	euc50	euc75	euc100	cos50	cos100

(表 13-結果比較)

從表 13 中，我們可見得除了仁寶之外，各個情境中最好的結果都落在取停用字與集合的方法。單純以此實驗結果推測，可以發現到於橘子與鴻海這類型股價偏中高且波動較大的股票類型，多以歐基里德距離算法勝出；而仁寶這種股價與波動皆偏低，且預測成果平均誤差也不大的情況下，全部文章引用會為模型帶來更好的成果。至於中華電信及兆豐金波動較小，且較好的成果共同出現在餘弦距離算法，但因其在股價上的類型剛好相反，因此在此部分無法論斷波動較小的股價適合以餘弦距離算法表示。其中還有一個共通點:波動較小且股價偏低的，例如仁寶與兆豐金，都是全部文章引用會為模型帶來更好的成果，因此可以初步推論，以關聯度方法篩選文章對於波動較小的股價效用不大。

另外，可以發現曼哈頓距離似乎都沒有上榜，但是似乎無法直接否定曼哈頓距離算法，因為以下的統計:



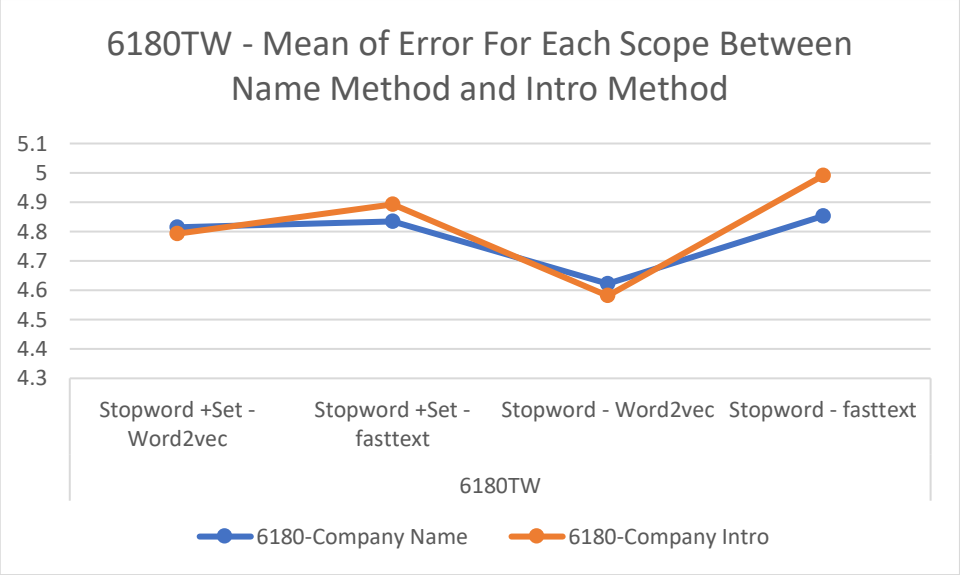
(圖 32 -各股票各個情境中前三最優秀的新聞提取方法)

根據圖 32，雖說曼哈頓距離並未出現於各個股票所有排列組合的 TOP1，但是於各個情境中(例如停用字+取集合+Word2vec；停用字+取集合+fastText；僅有停用字+Word2vec；僅有停用字+ fastText)，曼哈頓距離仍有佔有一席之地。其在每個股票中的四種情境中，拿到了 5 次的情境 TOP1，雖然在各個股票的所有情境的比較中並非是 TOP1，但不能推斷曼哈頓距離算法不適合，僅能推論是否因為實驗特性或是股票類型未有其表現的空間。

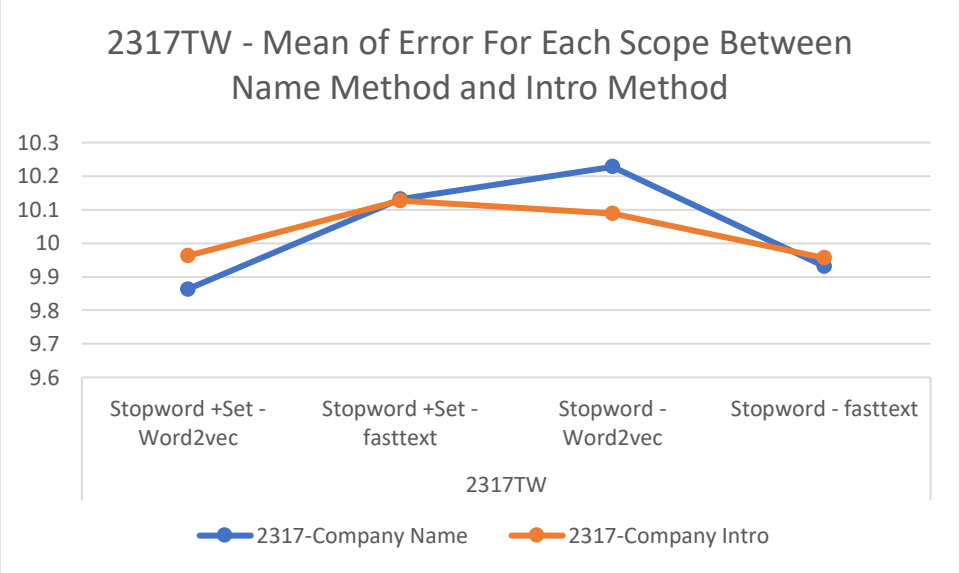
至此已介紹完兩種模式，分別以公司名稱及公司維基百科介紹做為中心文章的結果展示。但是你一定會好奇，這兩種孰優孰劣?結果差在哪裡?本研究將於下一小節對這兩種模式做比較及統整。

5-5 Compare Two Method

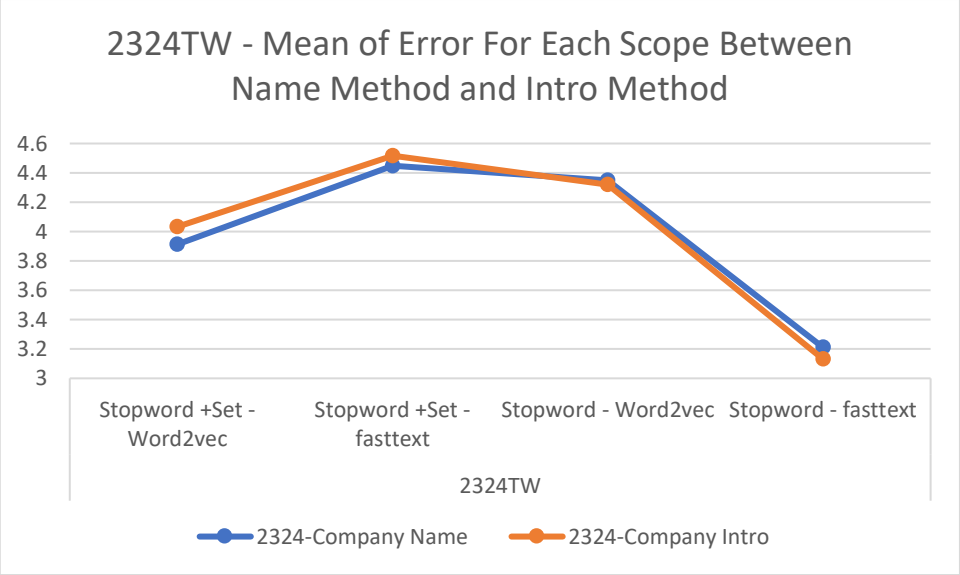
本小節將著重於比較以公司名稱作為中心點或是以公司 wiki 介紹作為中心點的兩者之差異。以下將比較 5-3 與 5-4 章節中各個表中，四種情境(下圖 33-37 的 X 軸)所有結果之平均誤差，並以公司名稱中心及公司 wiki 中心兩個模式作為基準，Name method 與 Company Name 皆代表以公司名稱作為中心；Intro Method 與 Company Intro 則表示以公司之 wiki 介紹做為中心。



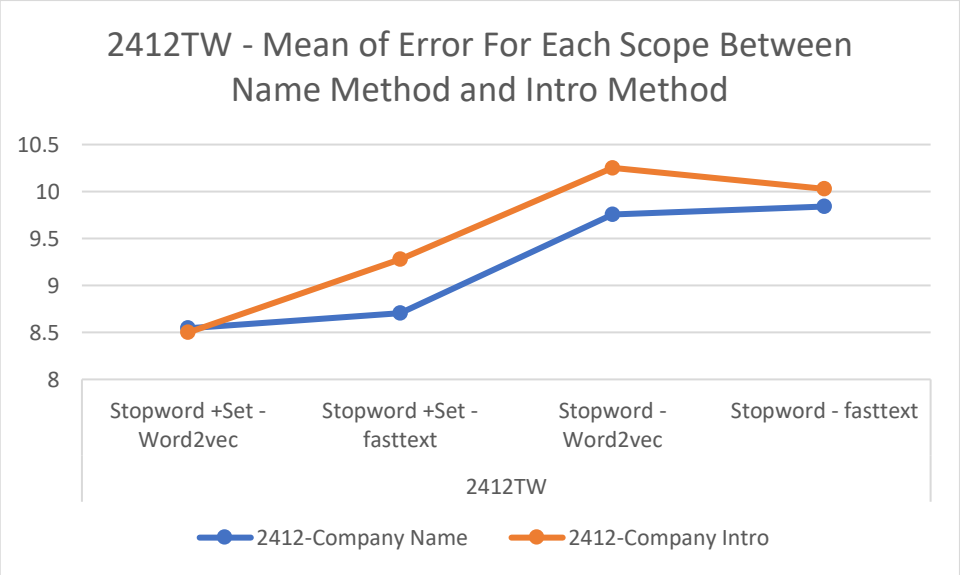
(圖 33 -遊戲橘子，各個所有情境預測結果平均誤差)



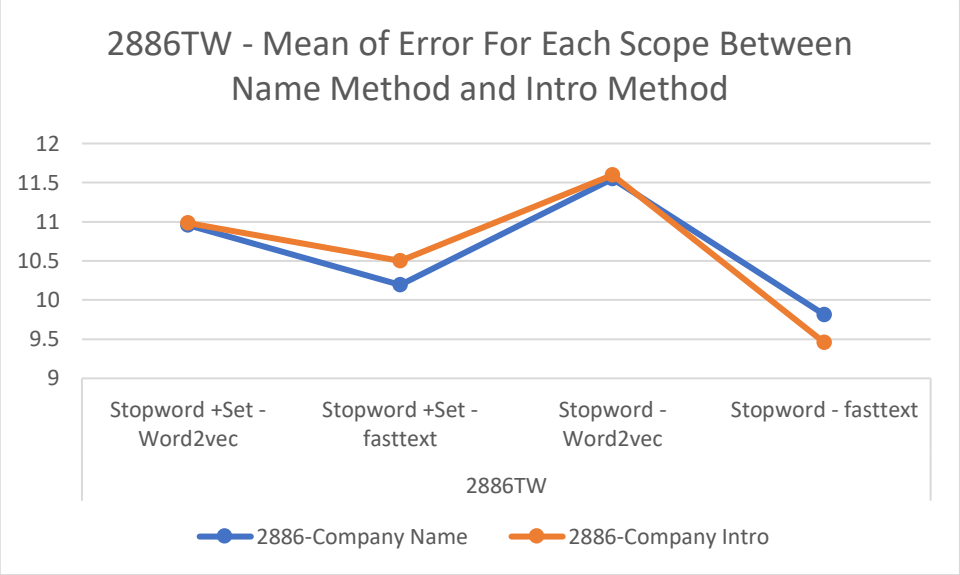
(圖 34 -鴻海，各個情境所有預測結果平均誤差)



(圖 35 -仁寶，各個情境預所有測結果平均誤差)



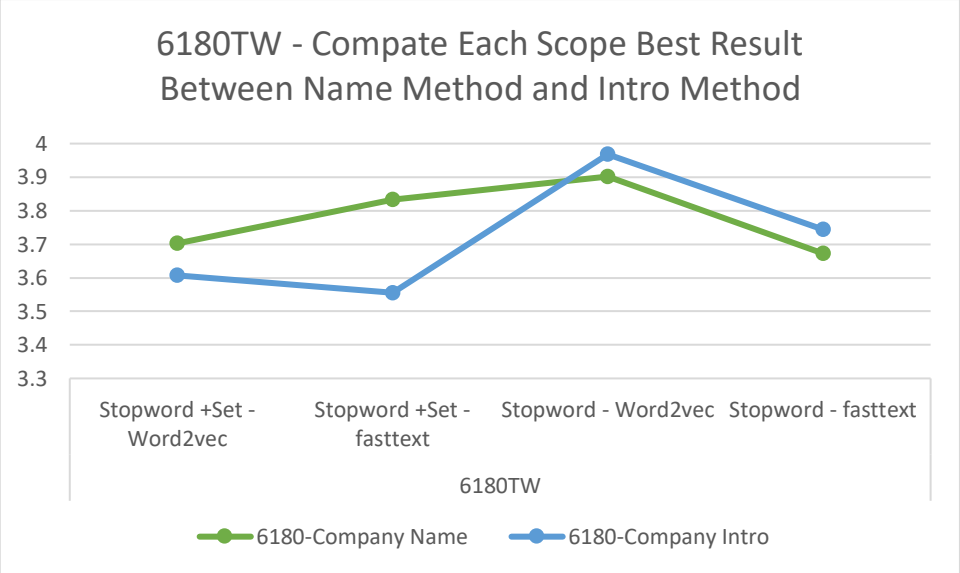
(圖 36 -中華電信，各個情境所有預測結果平均誤差)



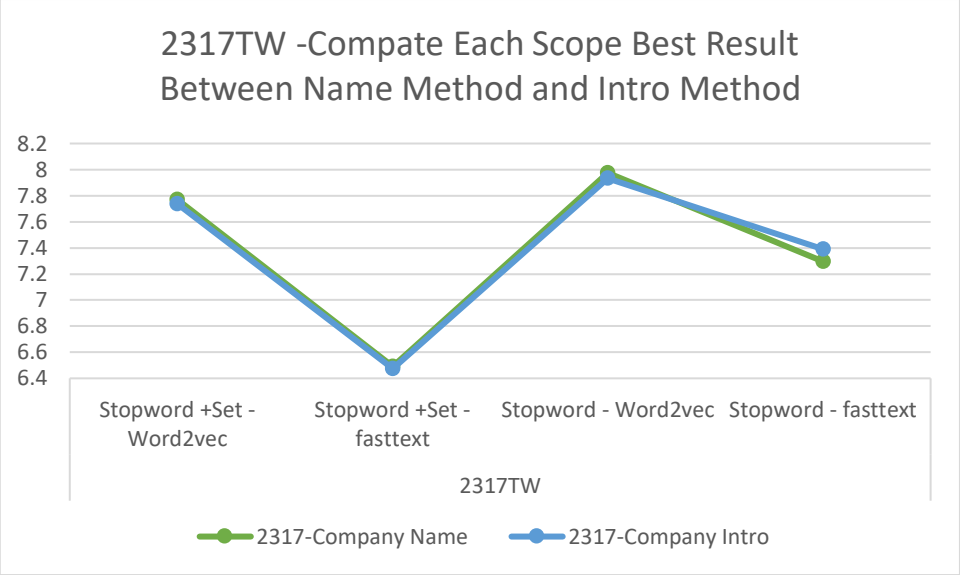
(圖 37 -兆豐金，各個情境所有預測結果平均誤差)

從圖 33-37 中，以公司名稱與以公司 wiki 介紹兩種方法互有優劣，並無明顯的一方較為勝出，僅有於中華電(2412)的情境中，皆以公司名稱作為中心的方法完全勝出。

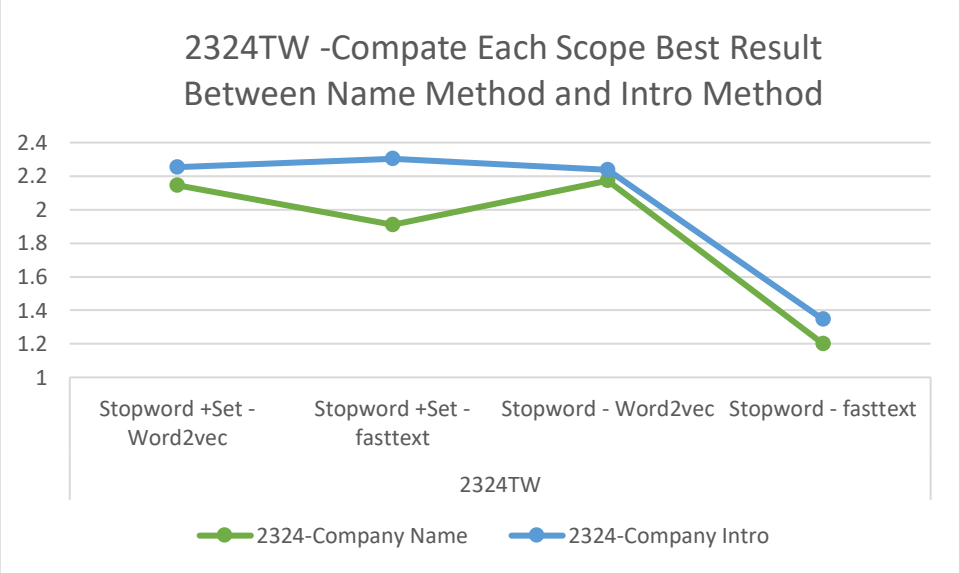
根據上述所有情境中所有平均誤差無法得到何者較好的結論，接著將範圍縮小，比較各個情境中表現最好的提取方式其之間的差異。



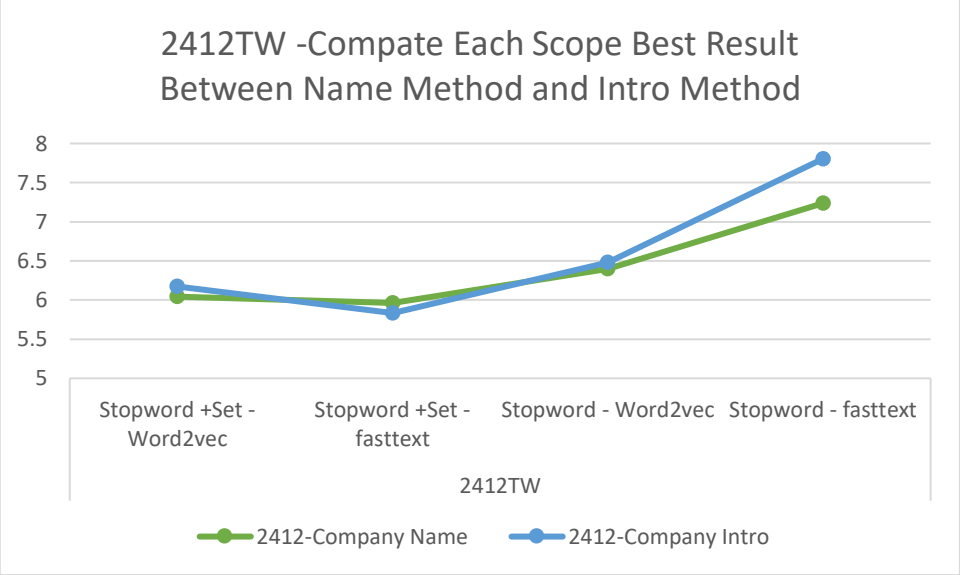
(圖 38 -遊戲橘子，各個情境中表現最好結果比較)



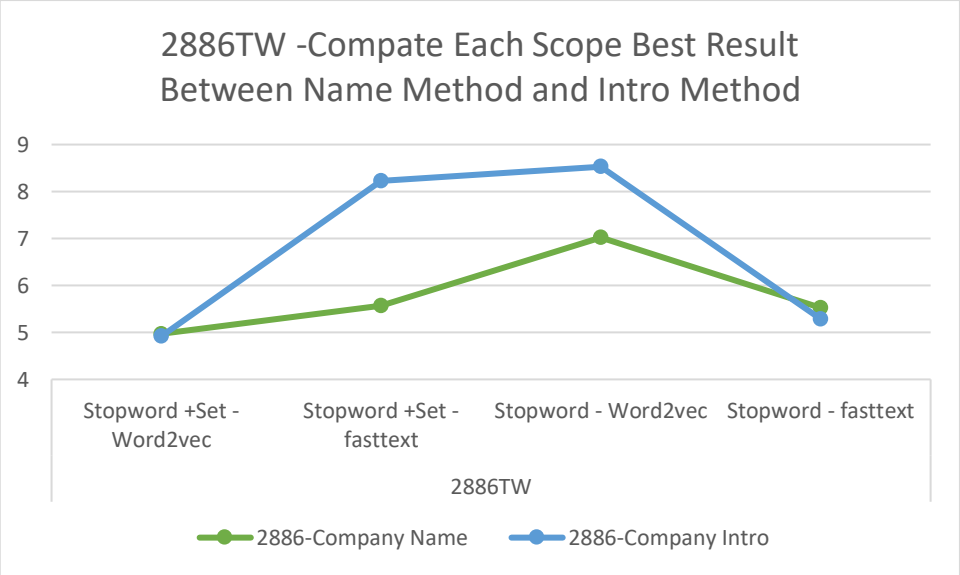
(圖 39 -鴻海，各個情境中表現最好結果比較)



(圖 40 -仁寶，各個情境中表現最好結果比較)



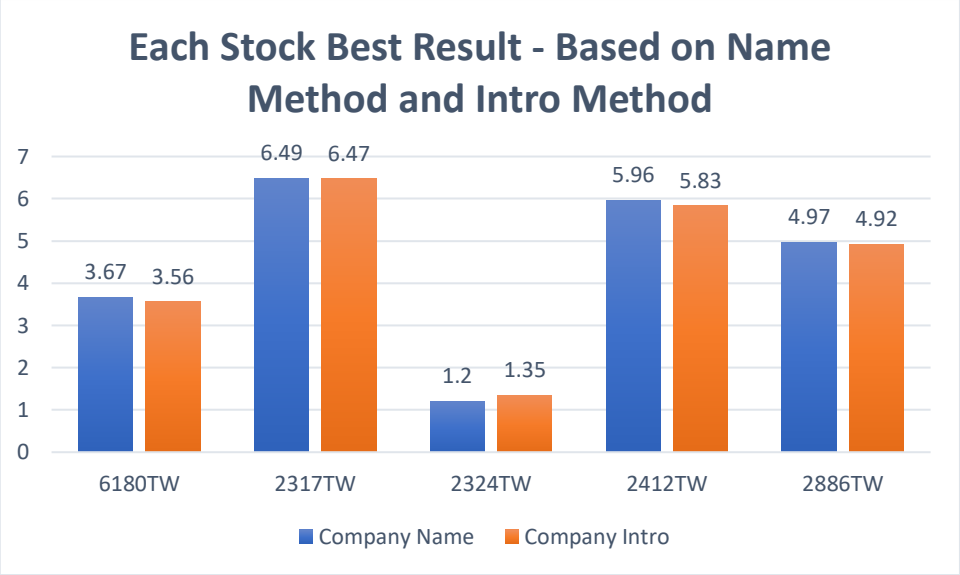
(圖 41 -中華電，各個情境中表現最好結果比較)



(圖 42 -兆豐金，各個情境中表現最好結果比較)

上圖 38-42 為四個情境中，表現最好的結果之比較。以遊戲橘子與中華電來說，兩者互有優劣，而鴻海則是兩者相當的貼近，兆豐金與仁寶則是皆以公司名稱作為中心點結果較好。從這樣的結果來說，同樣難以評定公司名稱作為中心或是公司 wiki 介紹何種是較好的方式。

有鑑於不論是四種情境(停用字+取集合+Word2vec；停用字+取集合+fastText；僅有停用字+Word2vec；僅有停用字+ fastText)的結果取平均，還是四種情境中取最好的來比較，似乎都無較好的結果，因此本研究再次縮小範圍，尋找各個股票中表現最好的預測結果之差異。



(圖 43 -各個股票中所有情境最好的結果比較)

上圖 43 中，本研究將範圍再次縮小，比較五種股票中，四個情境及 12 種文章提取標準中最好的結果並加以比較(值越小越好，代表預測與真實值差異越小)。從圖表中可見得單以表現最好的預測結果來說，除了仁寶之外，其餘股價在以公司 wiki 介紹做為中心的情況下結果較為出色。

Compare Table					
	遊戲橘子 (6180TW)	**鴻海 (2317TW)**	**仁寶 (2324TW)**	**中華電信 (2412TW)**	**兆豐金 (2886TW)**
Class1-Stock Price	股價中間	股價偏高	股價偏低	股價偏高	股價偏低
Class2-fluctuation	波動較大	波動較大	波動較小	波動較小	波動較小
Company Intro - Best Result-Method	**Stopword+Set (fastText)**	Stopword+Set (fastText)	Stopword (fastText)	Stopword+Set (fastText)	Stopword+Set (word2vec)
Company Intro - Best Result- News Extract	**euc50**	euc75	euc100	cos50	cos100
Company Name - Best Result-Method	**Stopword (fastText)**	Stopword+Set (fastText)	Stopword (fastText)	Stopword+Set (fastText)	Stopword+Set (word2vec)

Company Name -Best Result- News Extract	cos100	euc75	euc100	cos50	cos100

(表 14-結果比較)

表 14 則統整了以公司名稱作為中心及以公司 wiki 介紹做為中心最好的預測結果之組合，綠底處為其不一樣的地方。以遊戲橘子(6180)來說，可見得以公司名稱作為中心點時，最好的結果是僅使用停用字且使用 fastText 情況下的 cos100 結果最好，但是改為以公司簡介作為中心時則變成以停用字加上取集合，使用 fastText 情況下的 euc50 最好，且根據圖 43，以公司簡介作為中心的方法結果要更好一些。其餘的情況皆無改變，除了仁寶之外，大部分的情況都以公司簡介作為中心可以獲得更加精進的最佳結果。

6. Conclusion

綜合第五章節所述，以各個股票最佳的結果來比較的話，以公司簡介作為文章中心的結果大部分會比僅以公司名稱作為文章中心的結果來的好，若是目標追求最理想的結果，使用公司簡介會較僅使用公司名稱較為容易得到更好的結果。

就實驗過程以來，可以發現如仁寶或兆豐金這一類股價偏低且波動也較小的股票，都以所有文章全部使用會有最好的結果，換句話說關聯式新聞提取方法對於股價低且波動小的股票的效果並不顯著。但例如遊戲橘子、鴻海、 中華電信等來說使用關聯式的方式篩選文章，能得到更好的預測效果，例如說遊戲橘子以歐基里德方法計算距離，且取前 50%最具關聯的文章最好；鴻海也同樣以歐基里德距離方法為好，且其取 75%最具關聯的文章最好；而中華電信則以餘弦相關性計算且取前 50%最具關聯的文章為好。

此次實驗結果尚無直接找到在何種類型的股票適用於何種距離計算方法、應取多少文章之規律，也或許所謂的規律在難以掌控的股票市場中根本不存在，但根據實驗觀察可以假設如遊戲橘子與鴻海此類股價中偏高，且波動大的股票來說，歐基里德距離是較好的計算距離方法；而中華電信此種波動較小的股票則以餘弦距離方法為好，但波動小的股票也有例外，如仁寶則是以歐基里德距離最好。上述最好的結果中都無曼哈頓距離的身影，但無法直接的否定曼哈頓距離，例如圖 21 及圖 32，可見得曼哈頓距離在四種情境中，也佔有一席之地，只是整體來說最佳的結果並無曼哈頓距離的身影。至於處理文字的方法，則可以發現大部分的情況下，以停用字加上取集合的方法，且使用 fastText 作為 Word Embedding model 容易得到更好的結果。

綜合來說，經過本次研究發現波動大、股價偏高的股票較適合歐基理德距

離算法；波動小的股票則較適合餘弦距離算法，且針對文章應提取數量來說，除波動非常小之股價較不適用於此種新聞文章篩選方法外，大部分的文章經過僅取前最具關連的 50%或 75%往往能取得更好的成果。另外本研究也發現，以股票 wiki 介紹做為中心的情境來說，取集合的方法，且使用 fastText 做為 word embedding model 也較容易獲得更好的結果

未來的目標將會朝向使用更多實驗樣本驗證各個結果是否一致外，也必須要設計更多新聞資料處理方法來做實驗，例如我們可以對於詞彙向量有不同的處理，又或許，我們可以設計完全不一樣的市場文章篩選方法，做為與此研究方法之比較。