

# 基於長短期記憶神經網路 以新聞事件進行股價預測

**News Event-Driven Stock Prediction  
Based on Long Short-Term Memory Network**

資管碩二 鄒雅雯

指導教授：曹承礎 博士

# Agenda

1

緒論

2

文獻探討

3

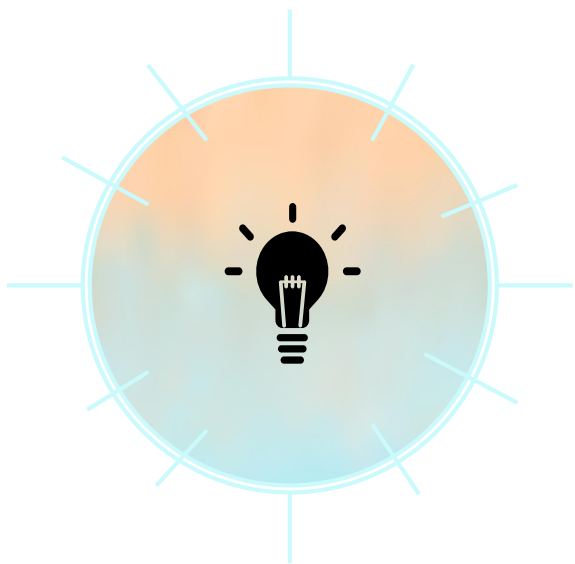
研究方法

4

研究結果

5

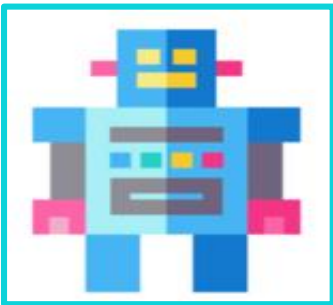
總結與未來研究方向



# 緒論

研究背景與動機 / 研究流程

# 研究背景 與動機



## 金融科技 機器人理財

若能直接對股價進行預測，尋找適合的買點及賣點，以提供機器人理財更精確的資訊，推播給客戶更完善、適合的投資組合。



## 台美連動關係


台灣許多主要的產業均是美國科技產業的代工，因此台灣的股市與美國股市有很大的連動關係



## 消息面

### Fama(1965)提出：金融市場為資訊有效的

股價會反應出所有已知的訊息，而大眾主要獲取資訊的來源即為新聞及報章雜誌，因此新聞發布之事件將影響股市的波動



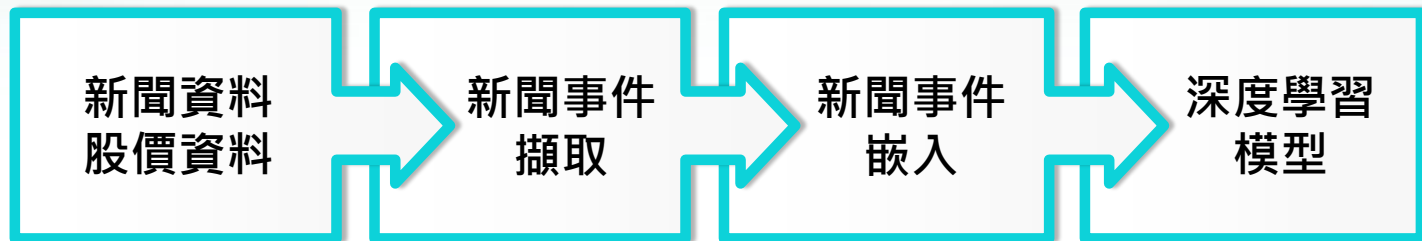
以美國之新聞資料

**路透社與華爾街日報**

對**台灣加權指數**

進行股價之預測

# 研究流程



01

美國路透社 & 華爾街日報  
Yahoo Finance 的台灣加權指數

02

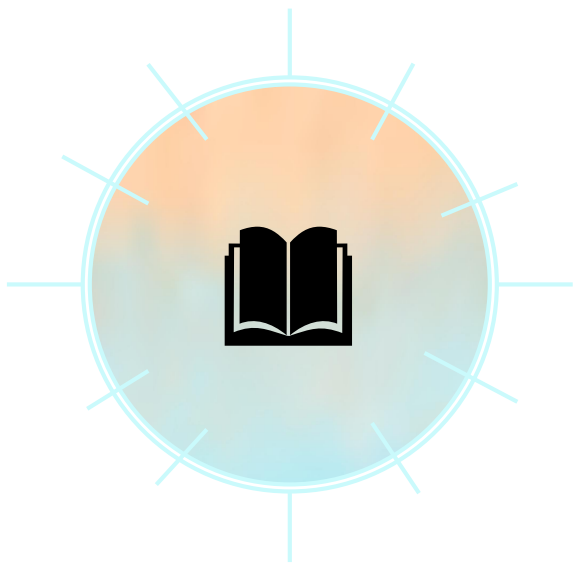
Open Information Extraction  
事件參與者 & 事件關係

03

事件嵌入 ( Event Embedding )  
AutoEncoder 模型

04

長短期記憶神經網路 (LSTM)  
時間序列資料



# 文獻探討

股價預測 / Open Information Extraction

# 股價預測

## 1. 股價預測類型

- Classification：二分類(漲/跌)、三分類(漲/跌/持平) → 準確率
- Regression：股價值的預測 → MSE

## 2. 股價預測資料來源

- 基本面：企業的基本資料，包含財務報表和非財務上的資訊，ex. Ou, J. A. (1989)
- 技術面：價量資料、技術指標，ex. Jing Zhang (2018)
- 消息面：財經新聞、企業新聞或社群網路中的文章等，ex. Bing, L. (2014)
- 結合2種以上的面向：ex. Deng, S. (2011)

## 3. 股價預測常用模型

- 統計的迴歸模型
- 機器學習：ex. Ince, H. (2007)使用支援向量機器( Support Vector Machine, SVM )
- 深度學習：ex. Akita, R. (2016)使用LSTM、Ding, X. (2015)使用CNN



# Open Information Extraction

## 1. 資訊擷取( Information Extraction )

- 從非結構化的資料中擷取出特定的訊息，轉化為結構化的表示方式( Jurafsky and Martin, 2009 )
- Ex.命名實體辨識( Named entity recognition )
- 主要在處理比較小，且已事先界定的問題上，意即要擷取的事件關係為事先界定的，而其使用的資料也是在較小且同質性高的資料集上

## 2. 開放資訊擷取( Open Information Extraction, Open IE )

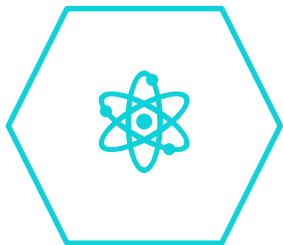
- 不受限於事先界定的事件關係，也不受限於資料集的領域，即使資料異質性很高，為跨領域的資料，Open IE亦能將事件關係提取出來
- (arg1; rel; arg2) , arg1 / arg2 : 事件參與者 ; rel : 事件關係
- She took the midnight train going anywhere. → (she; took; midnight train)

# Open IE系統

## Learning-based Systems

從訓練資料中自動學出句子的規則

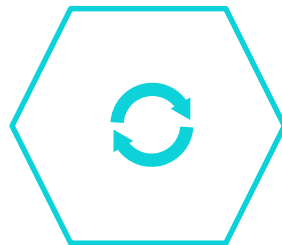
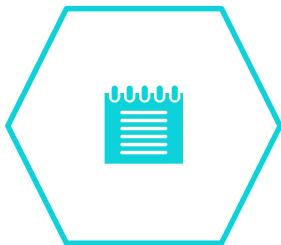
TEXTRUNNER (Banko et al., 2007)  
OLLIE (Mausam et al., 2012)



## Clause-based System

將複雜的句子重新建構

ClauseIE (Del Corro and Gemulla, 2013)  
Stanford Open IE (Angeli et al., 2015)



## Rule-based System

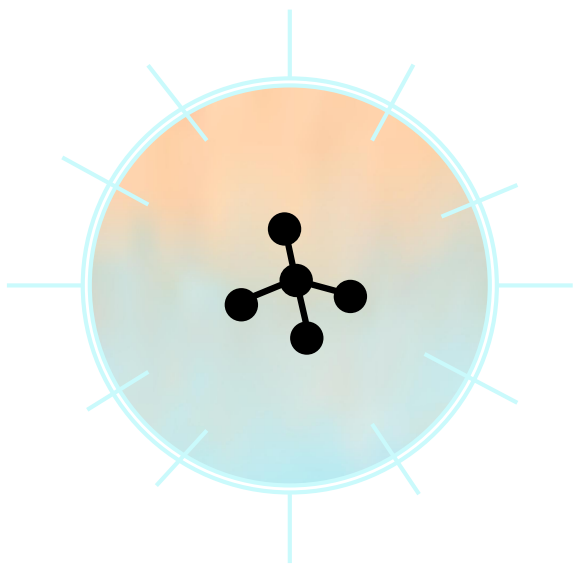
建立手工的提取規則

REVERB (Fader et al., 2011)  
PROPS (Stanovsky et al., 2016)

## System Capturing Inter-Proposition Relationships

取出事件間的關聯，呈現完整的全貌

OLLIE (Mausam et al., 2012)  
NESTIE (Bhutani et al., 2016)



# 研究方法

資料集 / 事件擷取 / 事件嵌入 / 深度學習模型

# 資料集說明



## 新聞資料

路透社 / 華爾街日報

財經類 & 政治類

2014/01/01-2019/11/30

共81,142篇

新聞標題 / 日期 /

新聞來源 / 主要內文



## 股價資料

Yahoo Finance

台灣加權指數(^TWII)

2014/01/01-2019/11/30

共1,463筆

日期 / 開高低收 /

調整後收盤 / 成交量

# 事件擷取

欲解決現行Open IE系統之不足

*In 1981, when staying in New York State, the pair developed the idea of running workshops for professional artists, which became the Triangle Arts Trust.*

## 重複性

相同的事件可能因有不同的修飾語，而產出內容重疊性高的事件，如此可能影響訓練資料的權重

### ClauseIE

("the pair", "developed", "the idea of running workshops for professional artists In 1981")  
("the pair", "developed", "the idea of running workshops for professional artists when staying in New York State")  
("the pair", "developed", "the idea of running workshops for professional artists")  
("professional artists", "became", "the Triangle Arts Trust")

### Open IE4

("the pair", "developed", "the idea of running workshops for professional artists")  
("the pair", "developed the idea of running workshops for professional artists In", ["1981"])  
("the pair", "developed the idea of running workshops for professional artists", ["when staying in New York State"])  
("professional artists", "became", "the Triangle Arts Trust")

## 擷取出之事件有缺漏

新聞的句子皆是較複雜且冗長的，包含了各式各樣的子句在當中，因此有些事件就被忽略了

# 事件擷取

欲解決現行Open IE系統之不足

*Labor is upset because many companies  
are using higher employee insurance  
premiums*

## ClauseIE

("Labor", "is", "upset because many companies are using higher employee insurance premiums")

("Labor", "is", "upset")

("many companies", "are using", "higher employee insurance premiums")

("many companies", "are using", "employee insurance premiums")

## Open IE4

("Labor", "is", "upset")

("Labor", "is upset because", "many companies are using higher employee insurance premiums")

("many companies", "are using", "higher employee insurance premiums")

## 重複性

相同的事件可能因有不同的修飾語，而產出內容重疊性高的事件，如此可能影響訓練資料的權重

## 擷取出之事件有缺漏

新聞的句子皆是較複雜且冗長的，包含了各式各樣的子句在當中，因此有些事件就被忽略了

## 沒有捕捉事件間的關聯性

若沒有將事件的上下文關係皆捕捉下來，則無法知道事件的全貌為何，因而傳達不正確的資訊

# 事件擷取

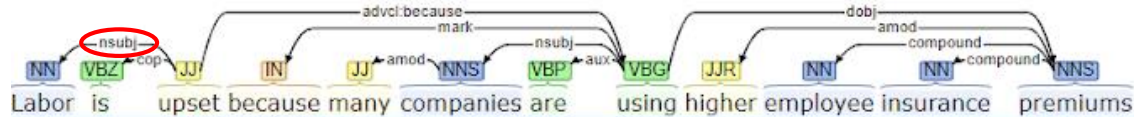
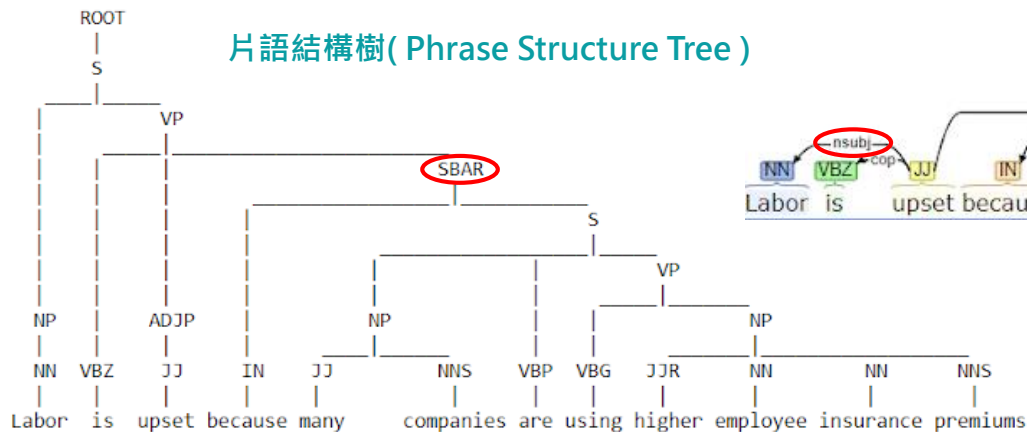
本篇論文所提出之Open IE

Clause-based

Divide-and-Conquer

Stanford NLP Group 2018

片語結構樹( Phrase Structure Tree )



依存關係樹( Dependency Tree )

# Divide-and-Conquer

## Divide

### 1. 連接詞單字 / 同位語

*Bell makes computer and building products.*



*Bell makes computer products.*

連接詞單字

*Bell, telecommunication company, makes building products.*



*Bell makes building products.*

*Bell is telecommunication company*

同位語

### 2. 關係子句 / 對等子句 / 副詞子句

*Bell, which is based in Los Angeles, makes building products.*



*Bell makes building products.  
which is based in Los Angeles*

關係子句

*Our team played hard but lost the game.*



*Our team played hard  
but lost the game.*

對等子句

*After finishing homework, I went out with my friends.*



*I went out with my friends.  
After finishing homework*

副詞子句



# Divide-and-Conquer

Conquer (arg1; rel; arg2)

## 1. 主詞

*Bell* makes computer products.

Diagram: A curved arrow labeled "nsubj" points from "Bell" to "makes".

一般句子

*Close the door, please.*

祈使句

*Bell, which is based in Los Angeles, makes building products.*

Diagram: A curved arrow labeled "acl:relcl" points from "Bell" to "which is based in Los Angeles".

關係子句

*After finishing homework, I went out with my friends.*

副詞子句  
對等子句

## 2. 動詞：因已經將長句子切割為可獨立成事件的短句子，因此短句子中僅含有一個事件關係

*The moon has not risen yet.* 取出連續動詞，修飾動詞之副詞(包含否定詞)後續將作為修飾語取出

## 3. 受詞：依據動詞的擷取結果

*Bell makes computer **products**.*

Diagram: A curved arrow labeled "dobj" points from "makes" to "products".

有受詞

*The moon has not risen yet.*

無受詞

# Divide-and-Conquer

(arg1 [arg1\_attr]; rel [rel\_attr]; arg2 [arg2\_attr])

## 取出修飾語

形容詞、介係詞、副詞

*In 1981, when staying in New York State, the pair developed the idea of running workshops for professional artists, which became the Triangle Arts Trust.*

{(the pair [In 1981]; developed [for professional artists]; the idea [of running workshops]);  
when (the pair [In 1981]; staying in; New York State);  
which (the idea [of running workshops]; became; the Triangle Arts Trust)}

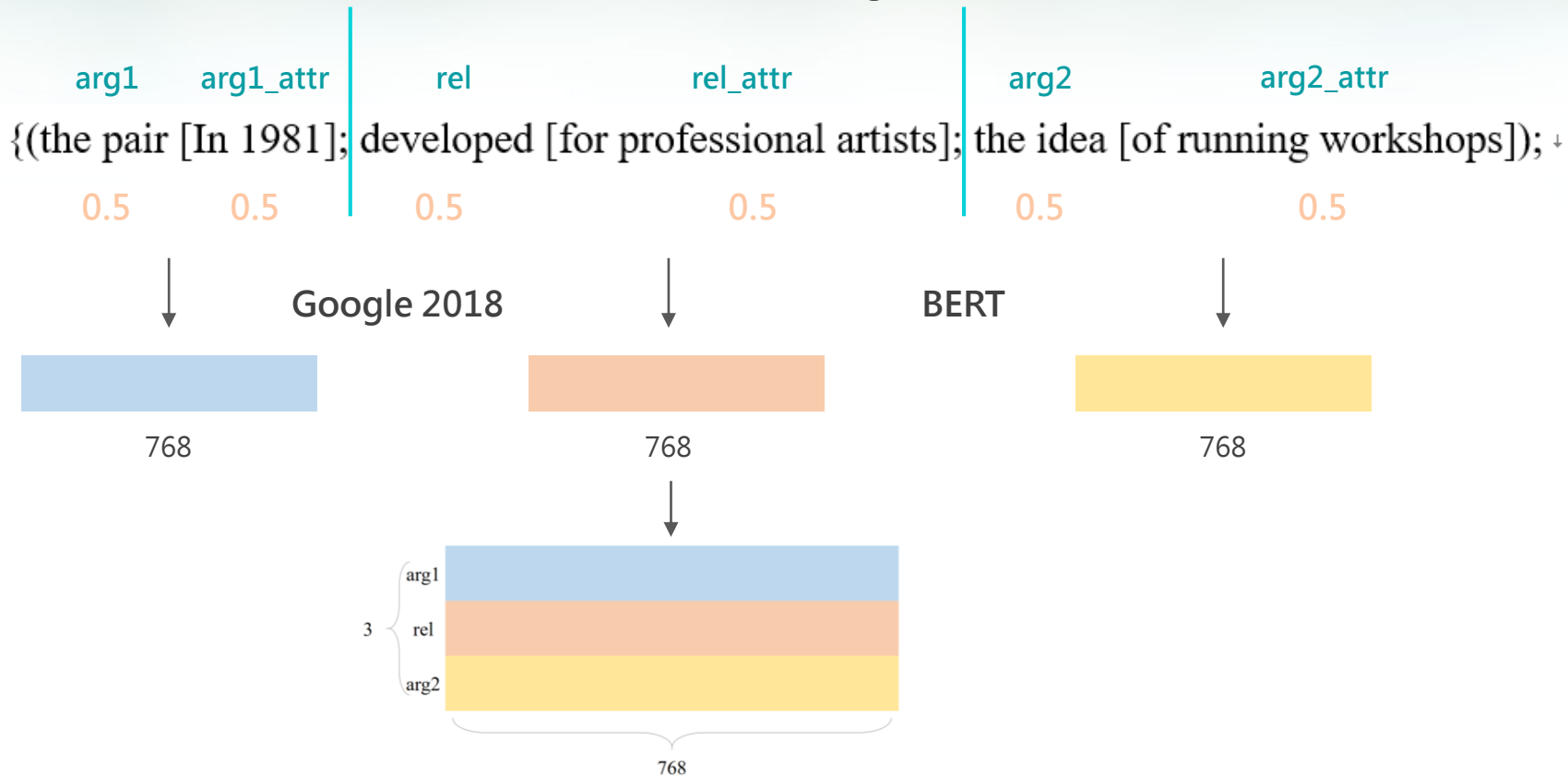
## 合併連接詞單字

*Bell makes computer and building products.*

{(Bell; makes; computer products); (Bell; makes; building products)}

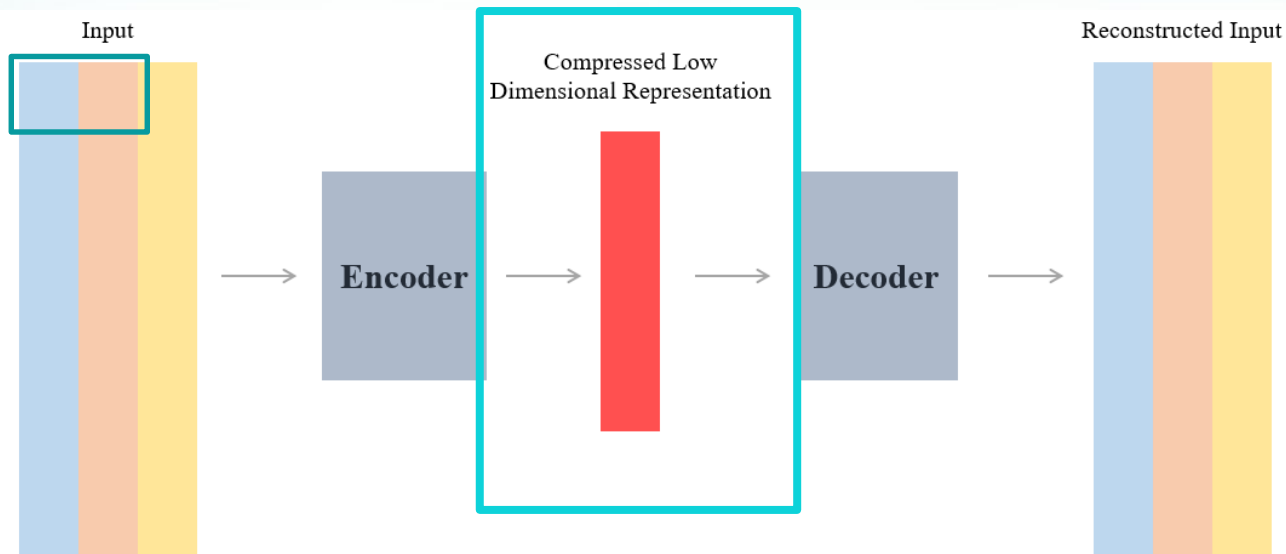
# 事件嵌入

Word Embedding



# 事件嵌入

## Event Embedding



Train : 路透社 & 華爾街日報 2006/10/20-2019/11/30 的新聞

# 深度學習模型

## LSTM

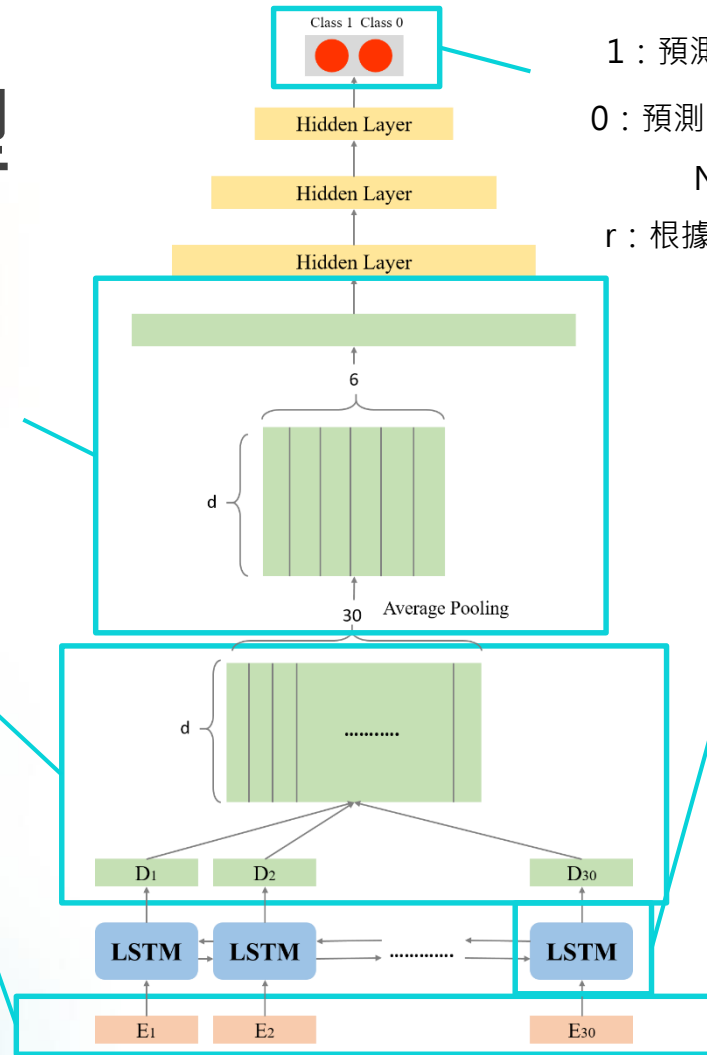
### 池化層 (Pooling Layer)

從每個維度、每5天的結果中取平均值

將每日的輸出結果取出  
得到 $R(d \times 30)$ 的矩陣

新聞事件的影響為遞減的

使用過往一個月的新聞事件作為輸入  
以天為單位，將每日發生的所有新聞事件進行平均，取得當天的事件向量(E)



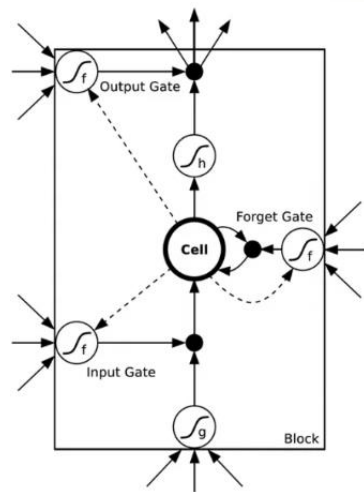
1 : 預測N日內台指報酬曾超過 $r$

0 : 預測N日內台指報酬未曾超過 $r$

N : 實驗結果中決定

$r$  : 根據樣本的正反比例進行調整

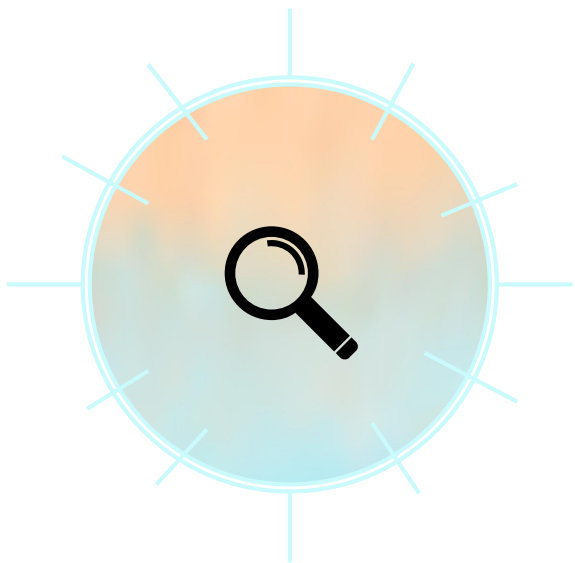
### 單個神經元



2 層隱藏層

每層隱藏層皆有800個神經元

雙向學習



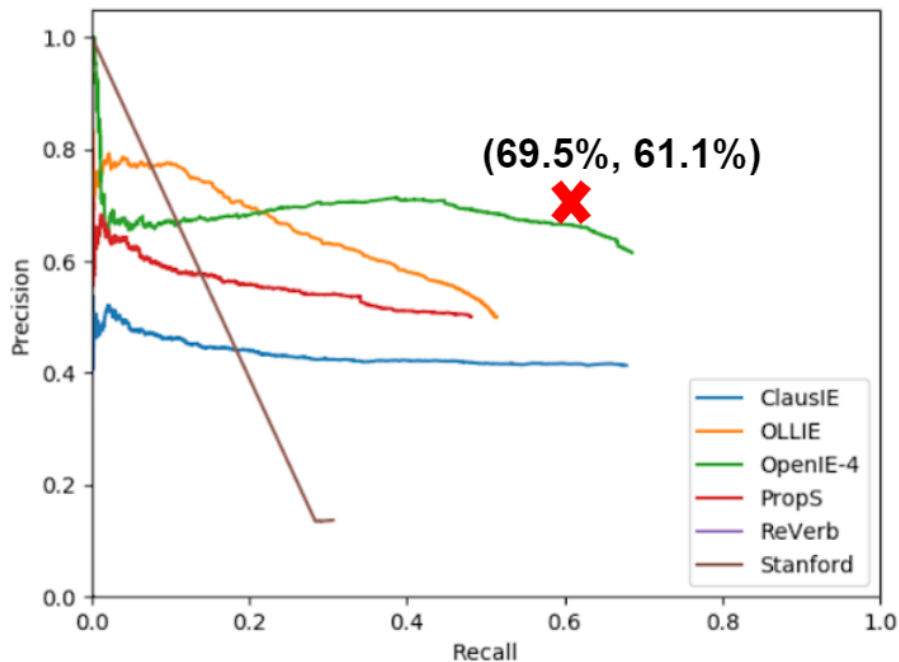
# 研究結果

Open IE 表現 / 實驗設定 / 驗證結果 /  
測試結果 / 牛市熊市測試結果

# 本篇研究提出之Open IE表現

Stanovsky, G. 2016年  
從QA-SRL的資料集轉換出  
的大型資料集Benchmark

本篇研究  
準確率為69.5%  
召回率為61.1%



# 實驗設定

- 資料來源：美國路透社( Reuters )及華爾街日報( The Wall Street Journal )的財經及政治類新聞
- 日期：2014/01/01 - 2019/11/30
- 模型輸入：只擷取新聞標題的事件
- 預測目標：N日內台指之報酬是否曾超過r





# 實驗設定

	文章數量	事件數量	時間區間
訓練資料(Training)	69,496	90,083	2014/01/01-2018/10/17
驗證資料(Validation)	5,210	6,926	2018/10/18-2019/05/06
測試資料(Testing)	6,436	8,850	2019/05/07-2019/11/30

**EB-LSTM**

Event Embedding + LSTM

**EB-XGBoost**

Event Embedding + XGBoost

**WB-LSTM**

Word Embedding + LSTM

**WB-XGBoost**

Word Embedding + XGBoost

# 驗證結果

不同預測天數 (N日內報酬曾超過r)

- 3日內報酬曾超過1.001視為1，若無則0
- 5日內報酬曾超過1.005視為1，若無則0
- 10日內報酬曾超過1.01視為1，若無則0

	Precision	Recall	F1
3日漲跌(1.001)	0.7005	0.7266	0.7077
5日漲跌(1.005)	0.6106	0.6719	0.6296
10日漲跌(1.01)	0.5280	0.6563	0.5745

# 驗證結果

## Proposed Model and Baseline Models

	Precision	Recall	F1
EB-LSTM	0.7005	0.7266	0.7077
WB-LSTM	0.6657	0.6719	0.6686
EB-XGBoost	0.6245	0.6953	0.6426
WB-XGBoost	0.6293	0.6328	0.6310

### EB > WB

事件擷取方式將語句中關鍵的參與者以及事件關係取出，事件嵌入將語句進行濃縮，學習之中的語義關係，萃取出事件之特徵向量

### LSTM > XGBoost

LSTM擅長處理時間序列資料，能從過往一個月之新聞事件中保留影響股價漲跌之特徵向量，且新聞事件發生的先後順序皆可能導致不同的結果

# 驗證結果

市場模擬

## 市場設定

1. 台灣加權指數為可交易的
2. 交易手續費為0.1425%，每次買進及賣出時皆會收取

## 交易策略

若預測三日內漲跌為 1



投資NT\$10,000買進台指，持有台指至多3日

倘若持有期間內之報酬超過1.001，則在當日使用收盤價將台指賣出

若無則持有至第三日，以第三日之收盤價賣出

每次要將台指賣出時  
皆會使用模型進行預測



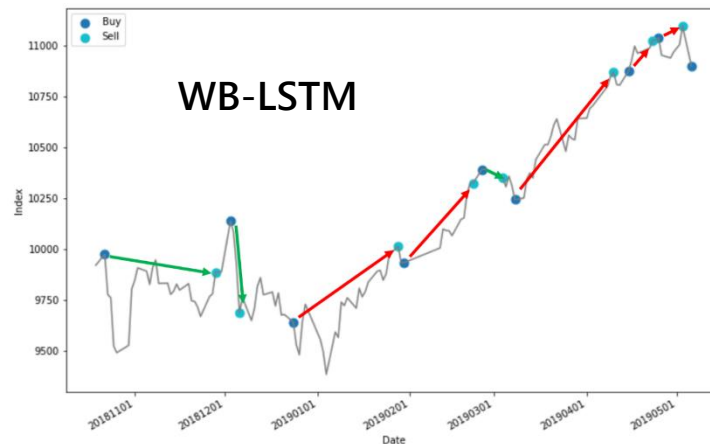
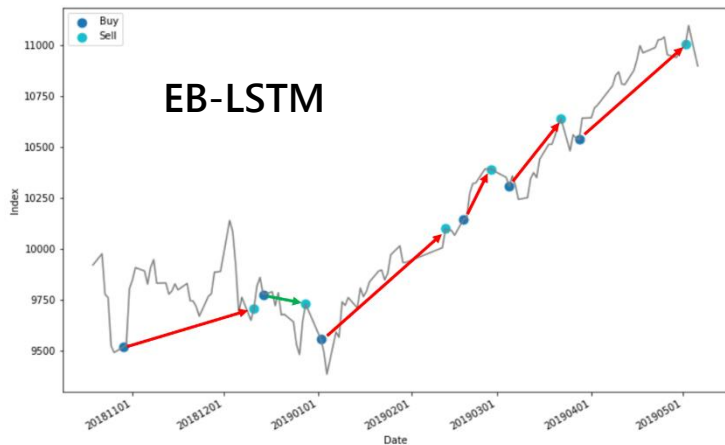
預測為1，繼續持有台指，觀察3日內之報酬

預測為0，將台指賣出，並結算進出場手續費

# 驗證結果

市場模擬

	Return		Return
EB-LSTM	1.1649	EB-XGBoost	1.0713
WB-LSTM	1.0730	WB-XGBoost	1.0422



# 測試結果

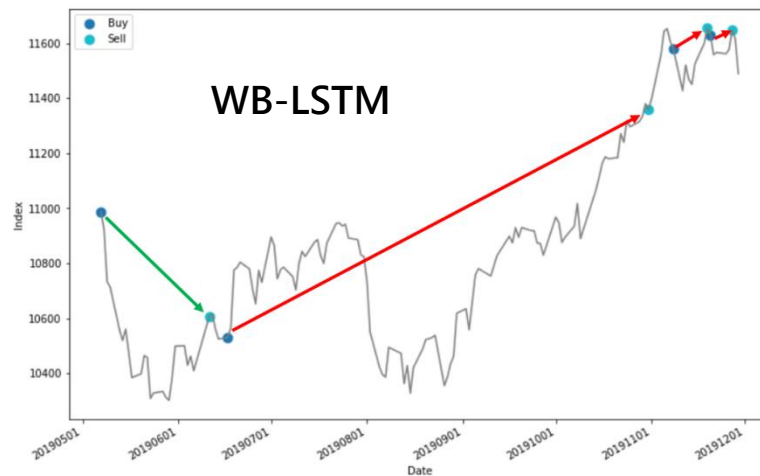
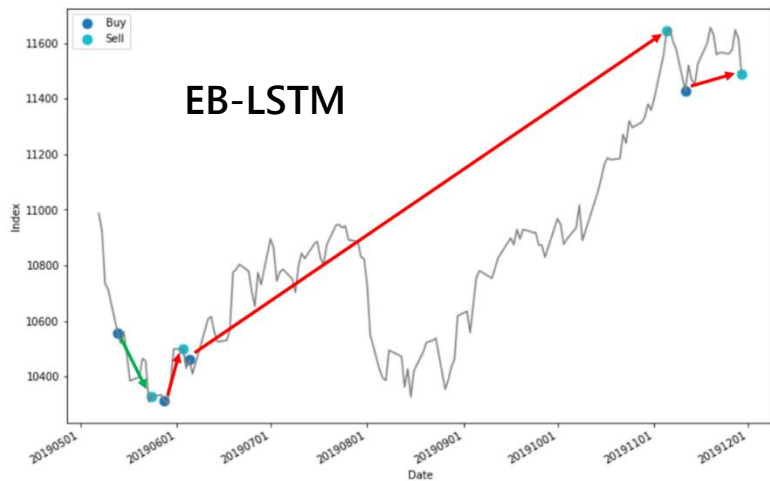
Proposed Model and Baseline Models

	Precision	Recall	F1
EB-LSTM	0.7134	0.6901	0.6250
WB-LSTM	0.6444	0.6620	0.5977
EB-XGBoost	0.6420	0.6620	0.6041
WB-XGBoost	0.6086	0.6408	0.5994

# 測試結果

市場模擬

	Return		Return
EB-LSTM	1.1022	EB-XGBoost	1.0420
WB-LSTM	1.0379	WB-XGBoost	1.0207



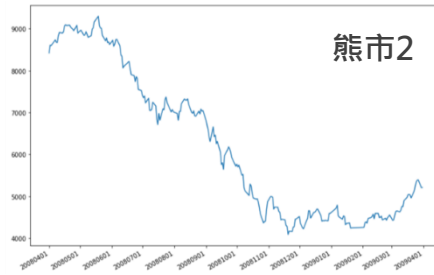
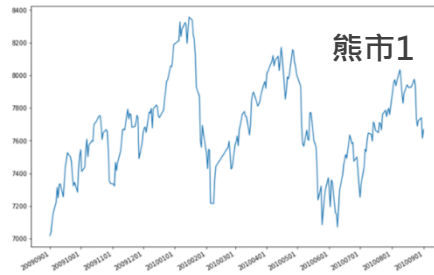
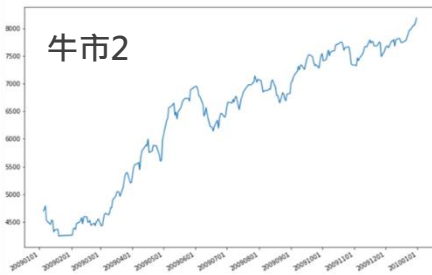
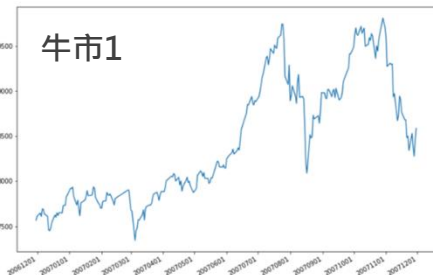
# 測試結果

## 牛市與熊市之測試結果

**牛市**：市場呈**上升**的趨勢，看漲情緒熱絡，股價**漲幅超過20%**

**熊市**：市場呈**下降**趨勢，投資人看跌，股價**跌幅超過20%**

時間區間			最大漲幅	時間區間			最大跌幅
牛市1	2006/11/30-2007/11/30		33.57%	熊市1	2009/09/01-2010/09/01		16%
牛市2	2009/01/01-2010/01/01		93%	熊市2	2008/04/01-2009/04/01		56%

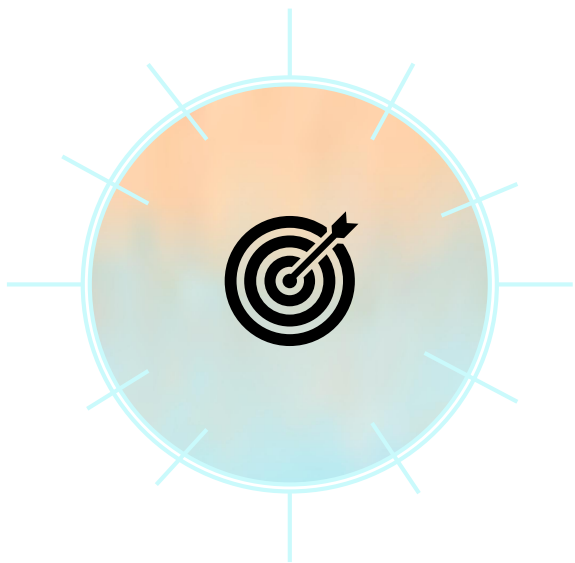




# 測試結果

牛市與熊市之測試結果

		Precision	Recall	F1	
牛市1	EB-LSTM	0.6431	0.7095	0.6413	65.16%
	WB-LSTM	0.5392	0.6390	0.5785	
牛市2	EB-LSTM	0.6287	0.7254	0.6619	56.83%
	WB-LSTM	0.6535	0.6598	0.6566	
熊市1	EB-LSTM	0.5782	0.6667	0.5983	56.83%
	WB-LSTM	0.5522	0.6382	0.5807	
熊市2	EB-LSTM	0.6139	0.6362	0.5382	56.83%
	WB-LSTM	0.5215	0.5656	0.5295	



# 總結

總結 / 未來研究方向

# 總結

1. 本篇研究提出新的的事件擷取方式，能更準確並完整地擷取出事件內容，亦能將事件間的關聯取出
2. **AutoEncoder**學習事件中的語義關聯，根據此模型能萃取出最具代表的事件向量，達到事件嵌入( Event Embedding )的效果
3. 考量新聞事件之影響並非短短一兩天，故使用過往一個月之新聞事件對股價漲跌進行預測，為要處理時間序列之資料，因此建立**長短期記憶神經網路(LSTM)**
4. 研究結果顯示，使用**事件嵌入( Event Embedding )**的方式較單純使用文字嵌入( Word Embedding )的方式更佳，而**LSTM模型**在捕捉有時序關係的新聞事件影響上較XGBoost模型表現佳
5. 在市場模擬方面，只需要使用最簡單的交易策略，根據模型之預測結果進行買賣便能達到不錯的報酬
6. 因近十年台指大多為牛市，因此模型的預測能力在**牛市**表現較熊市好

# 未來研究方向



## 擴展模型之資料來源

消息面

技術面

(價量資料、技術指標)

基本面

(企業產業資訊等)



## 修改深度學習模型

加入注意力機制

( Attention Mechanism )

為每天的事件賦予不同權重

辨別出影響股價重要的事件

並行的運算，加快模型速度



**Thank you**  
Q & A