



玉山金控 E.SUN FHC

×NTU

AI News Scoring System

智能新聞評分系統

D06943020 NTU EE 鄧傑方

B05602021 NTU BSE 連奕茹

B05701103 NTU BA 林韋丞

B06702064 NTU ACCT 林聖硯

Mentor : Prof. 蔡芸琿、詹益安(from E.Sun)



玉山金控
E.SUN FHC

Presentation Outline

- Overview
- Data Analysis
- Machine Learning Models
- Feature Analysis
- Conclusion & Future Works

The background is a soft-focus painting of a field of white daisies with yellow centers. A solid teal rectangular box is positioned in the upper left, containing the text 'Section I' in white. Below this box, the word 'Overview' is written in a large teal font on a white background.

Section I

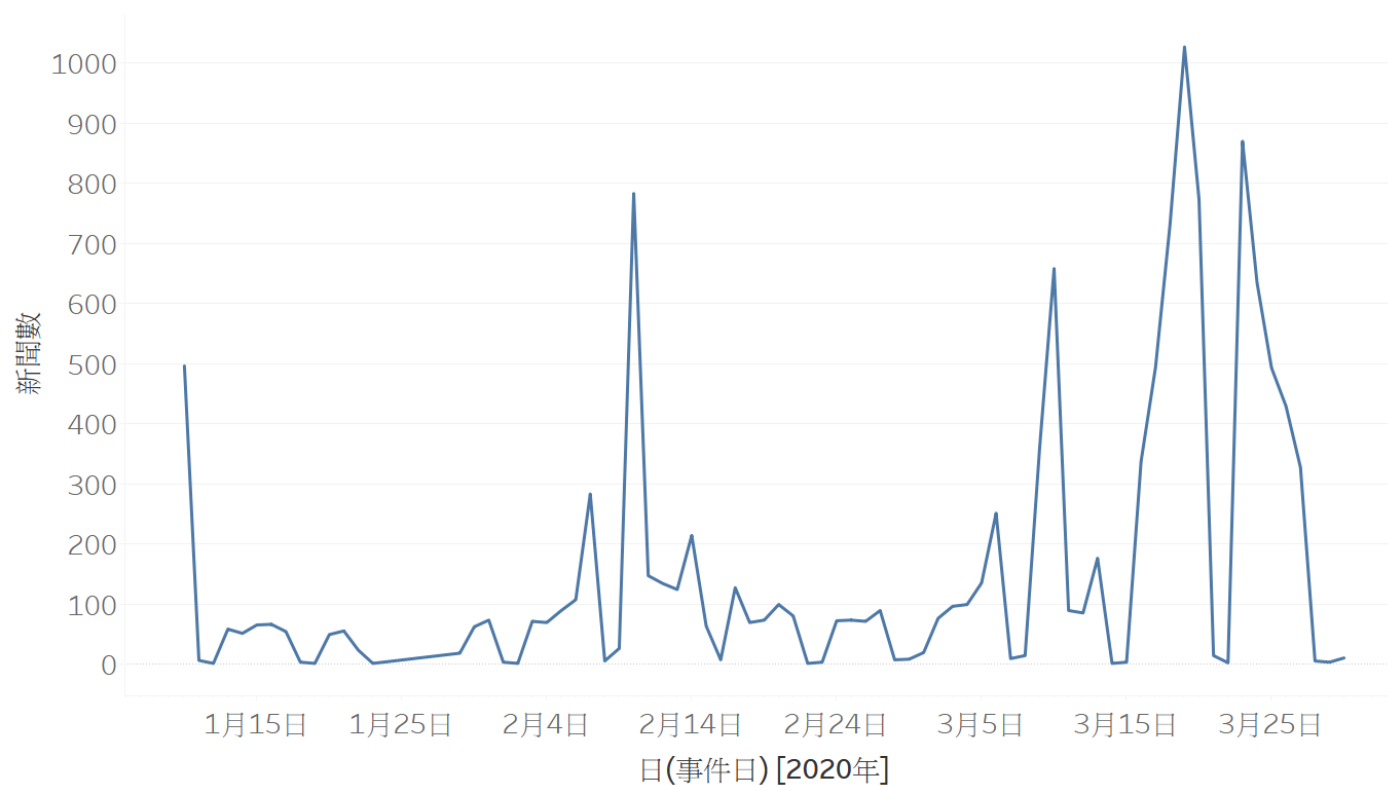
Overview

Project Introduction

- Consultants in TCRI judge financial news and rate them (+3 ~ -3)
- Pain Point
 - Heavy load of news everyday
 - Inconsistent judgement between experts
 - Important events can't be highlighted right away
- Target:
 - Using machine learning as a support system to help judge the news

Pain Point - Heavy load of news everyday

- Financial related events that happened from January 2020 to March 2020



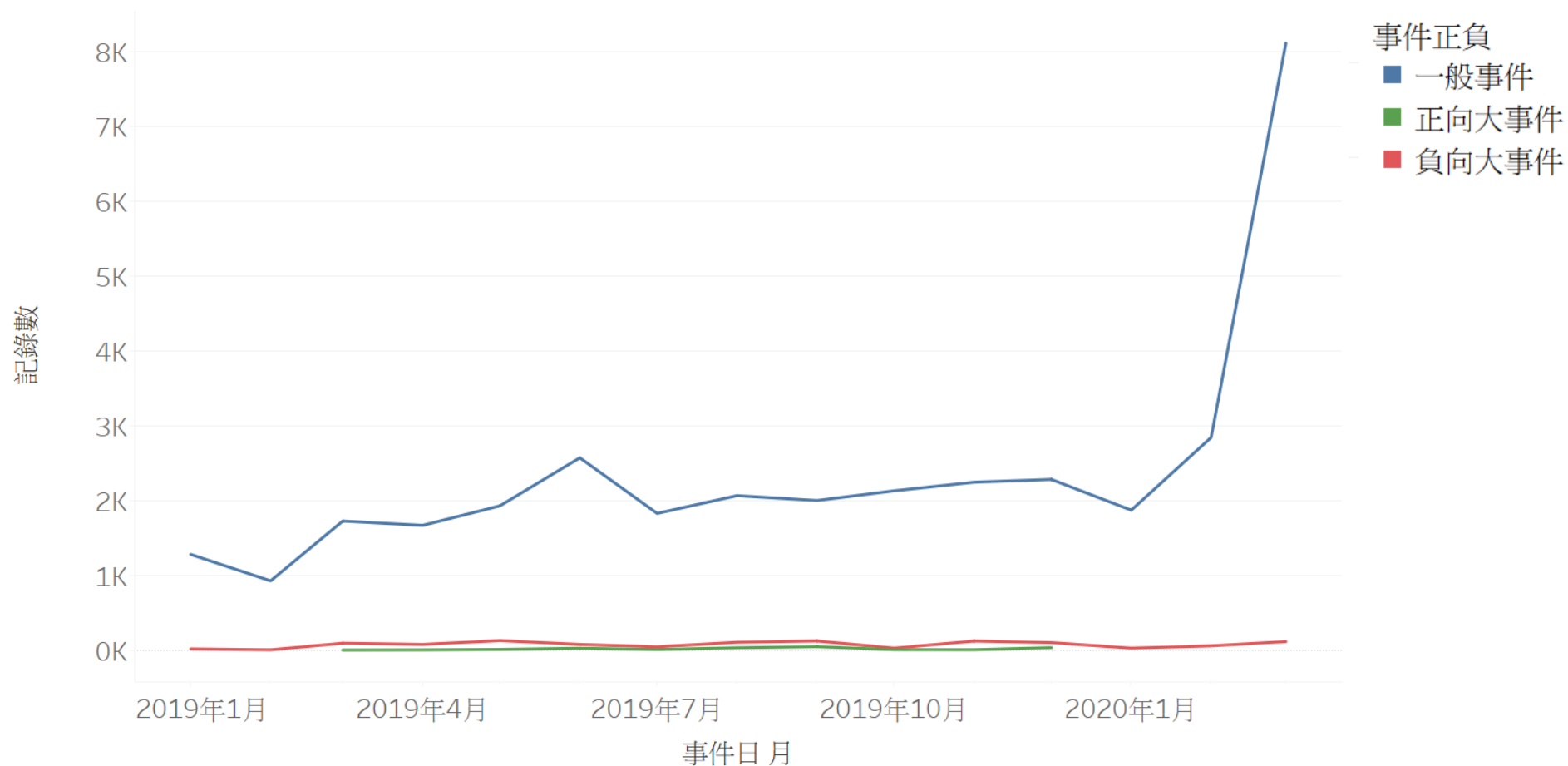
Pain Point - Inconsistent Judgement between experts

- Similar events was graded differently.
- Even same event was graded differently on the same day.

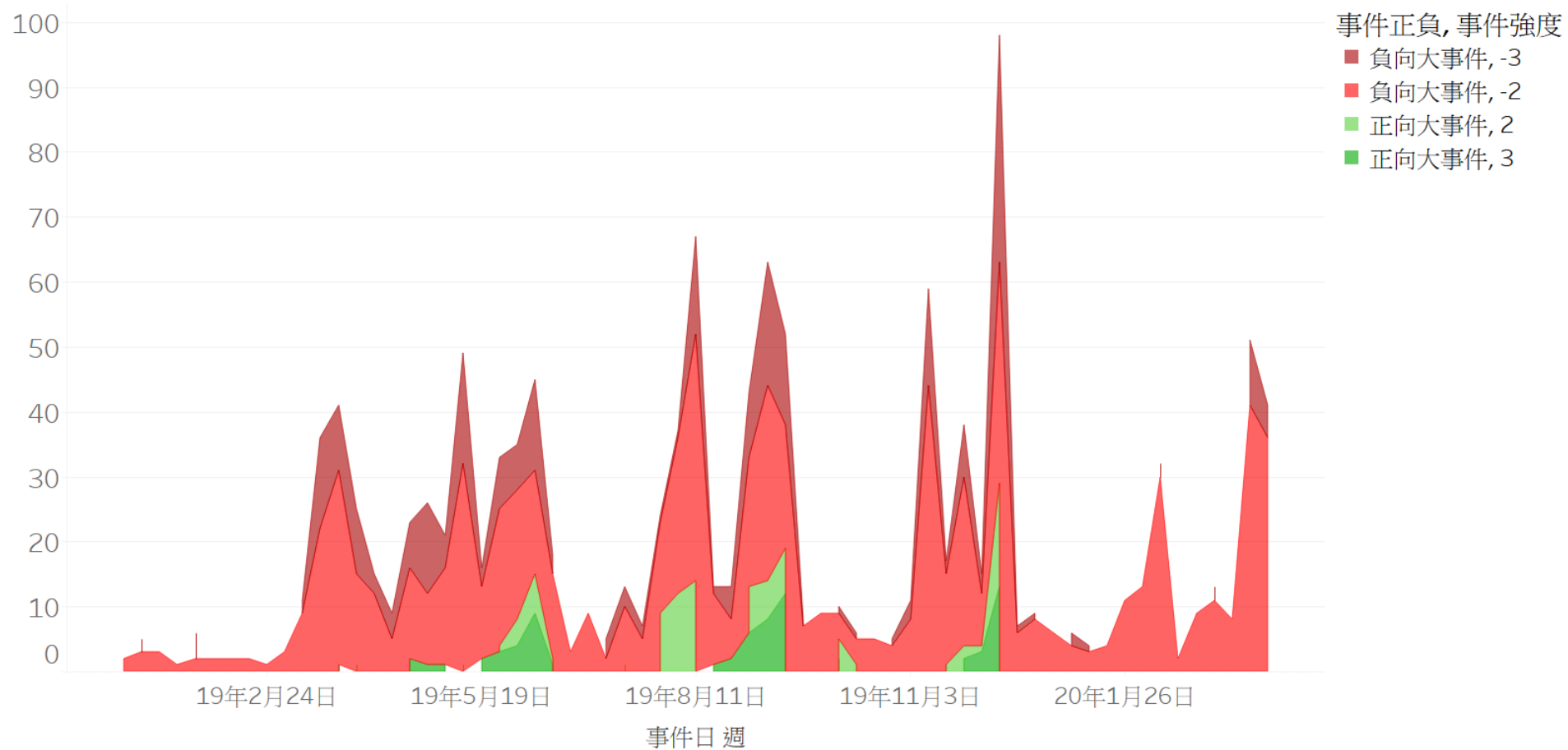
事件內容	事件強度
2019年11月累計營收655,726千元，年減32%。2019年11月單月營收90,853千元，年增32%。。	0
2019年12月累計營收726,769千元，年減29%。2019年12月單月營收71,043千元，年增29%。。	1
2019年09月累計營收6,323,164千元，年增51%。2019年09月單月營收408,266千元，年減23%。。	0
2019年10月累計營收6,645,678千元，年增42%。2019年10月單月營收322,514千元，年減35%。。	0
2019年11月累計營收7,125,647千元，年增35%。2019年11月單月營收479,969千元，年減17%。。	1
2019年10月累計營收361,638千元，年減31%。2019年10月單月營收72,822千元，年增58%。。	0
2019年12月累計營收453,215千元，年減24%。2019年12月單月營收33,933千元，年增35%。。	1

事件內容	公司簡稱	日(事件日)	事件強度
發言人[]動，由[]接任。。財務..	[]	2019年9月30日	-1
			0
		2019年10月1日	-1
			0

Pain Point – Important events can't be highlighted right away



Pain Point – Important events can't be highlighted right away



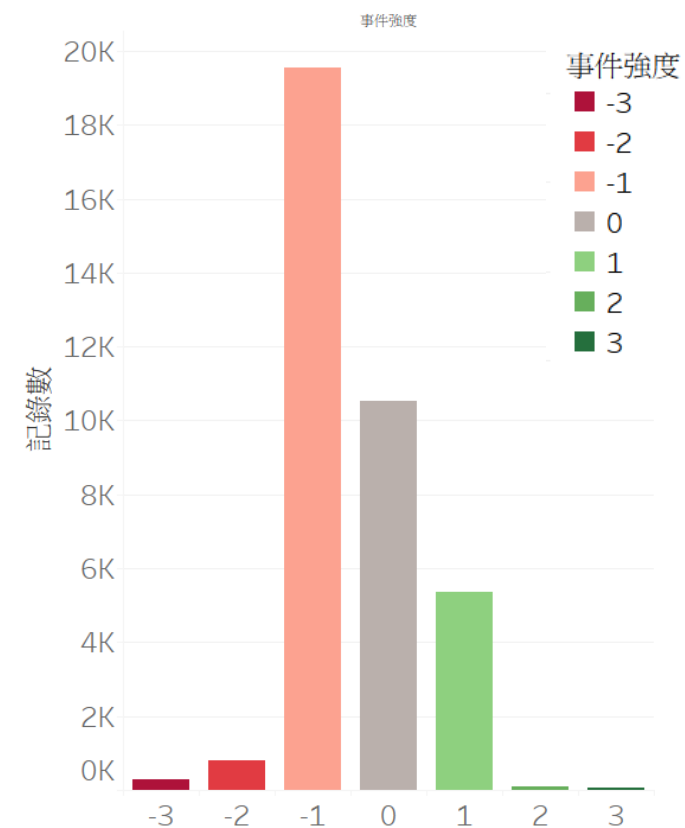
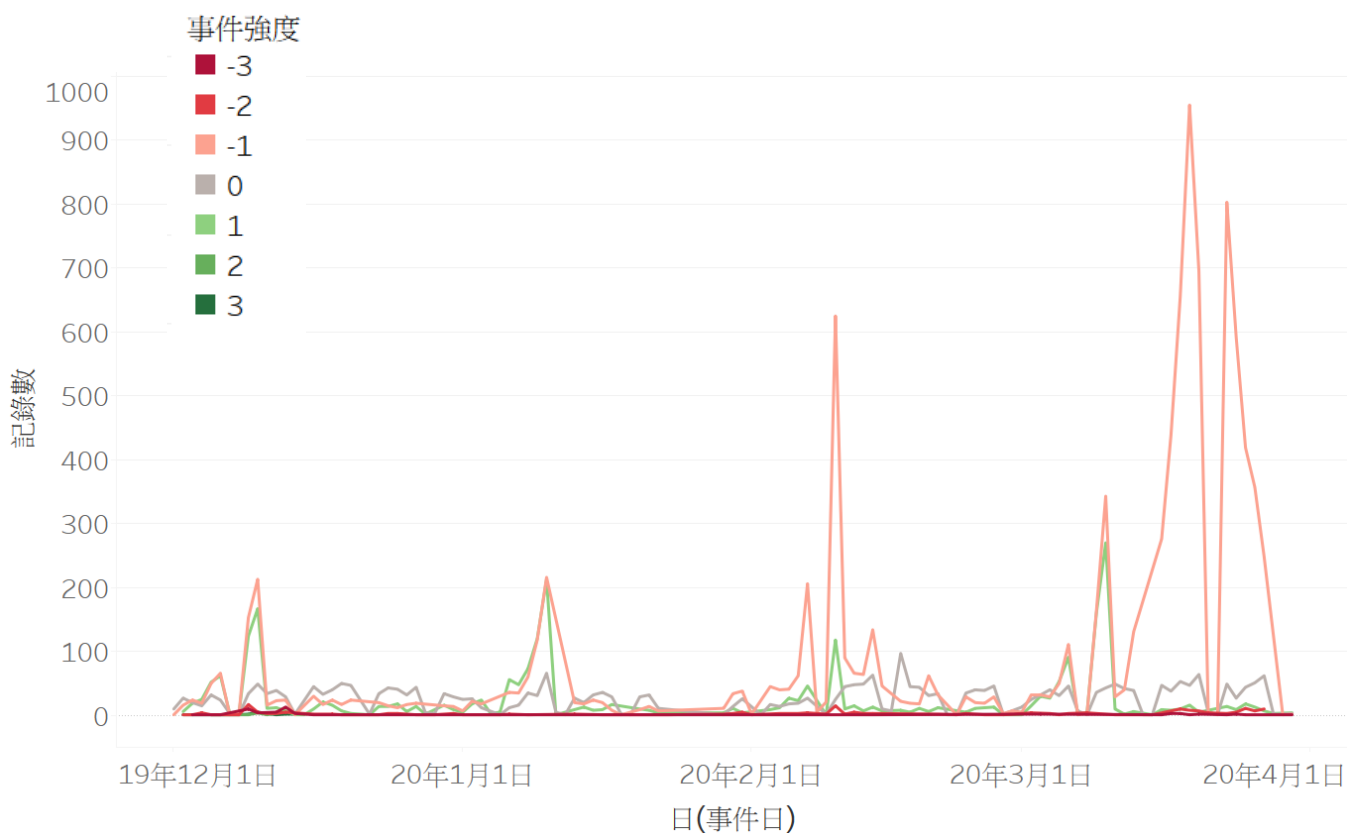
Section II

Data Analysis

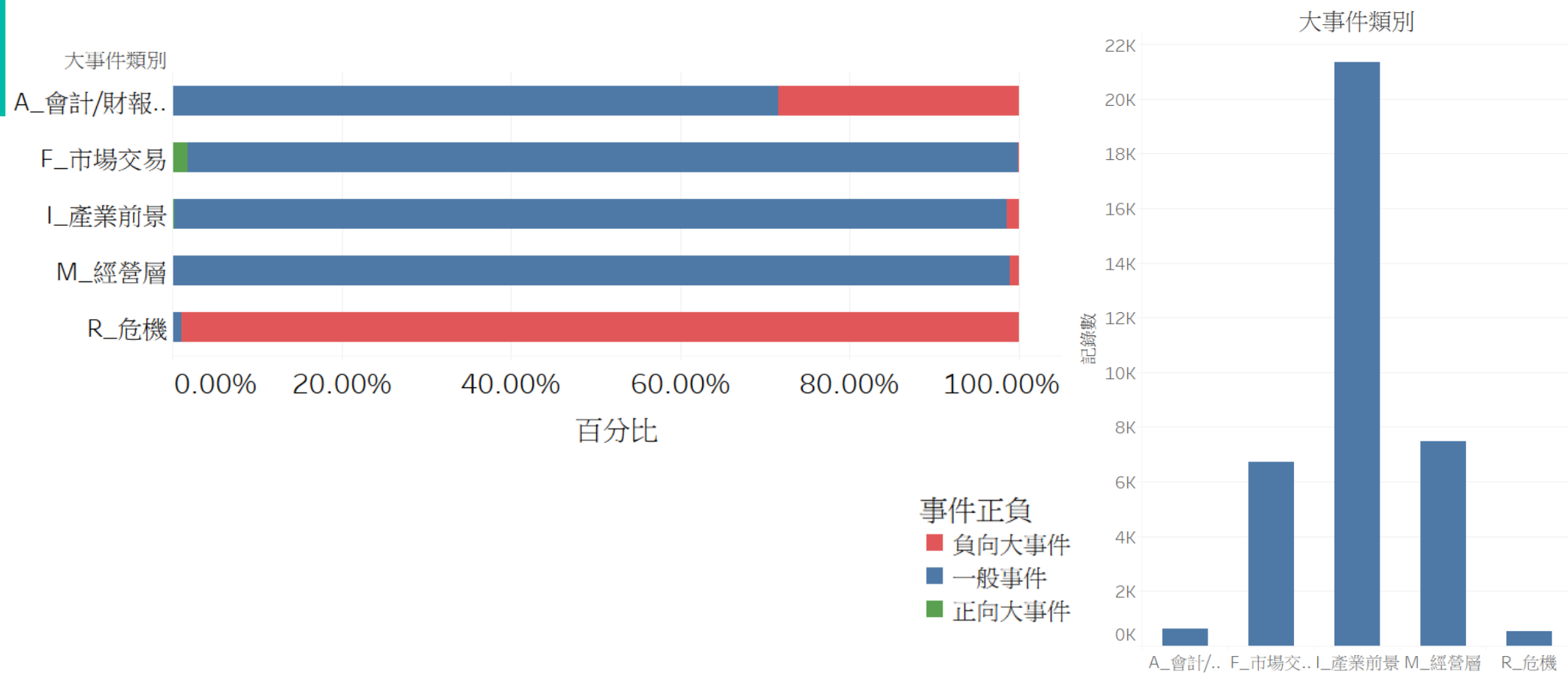
Data Set – Exploratory Data Analysis (EDA)

- Data description
 - 36,717 rows
 - From January 2019 to March 2020
- Influence of the event
 - -3 ~ +3
- Big event number: 5
 - Small event number: 99
- Corporation number: 1,980
- TCRI types: 1 to 9, C, D, Null
- Contents number: 35,338

EDA – Influence of the event



EDA – Big Event



TCRI Introduction

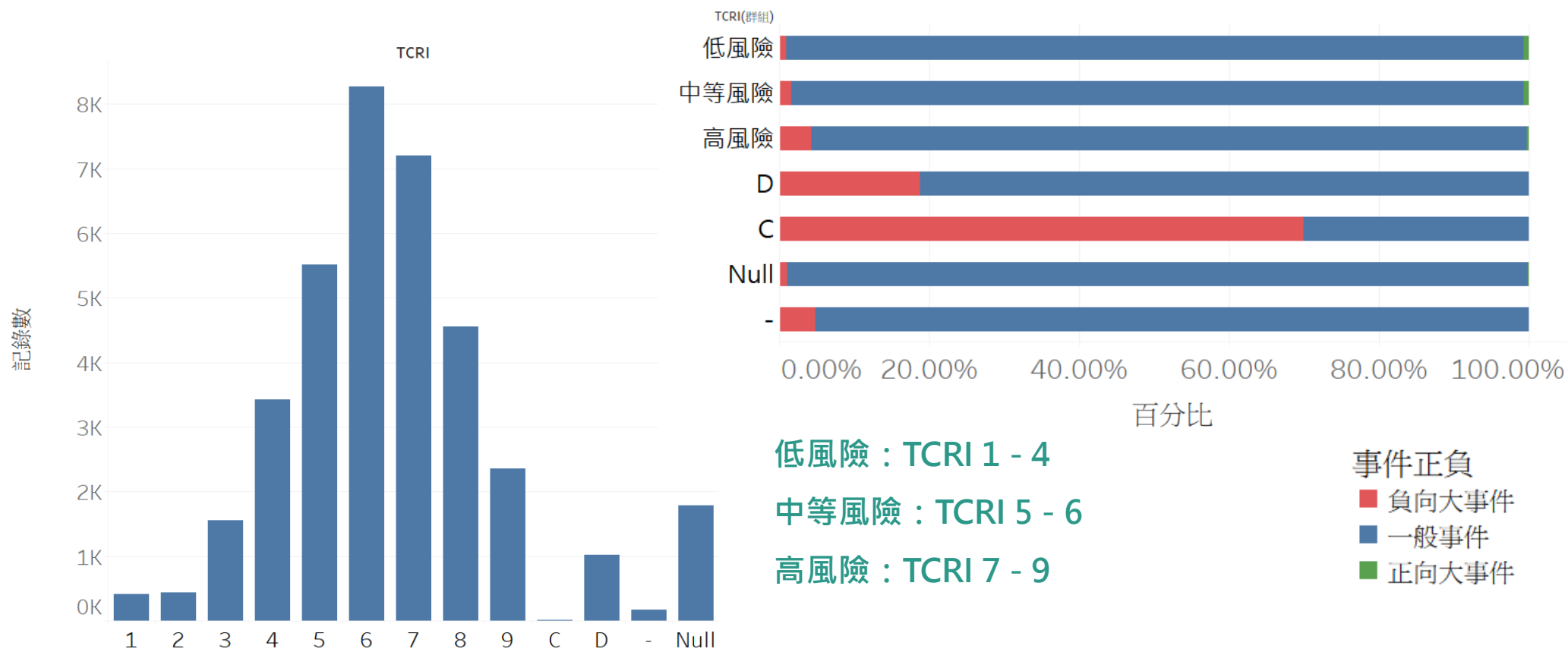
- What is TCRI:

- Experts judge most of the public companies except for some industry like financial industry or companies that are founded within four years.

- TCRI Ranking:

- 1~4 Cash Flow Lending
- 5~6 Cash Flow Lending/Asset Lending
- 7~9 Asset Lending Level
- D Financial Crisis
- C Didn't Disclose Financial Statement In Time

EDA – TCRI



Data Preprocessing – Segmentation System



JIEBA

System	CKIP	Jieba(結巴)
Developer	Academia Sinica	M.S. Student from China
Model	BiLSTM + CNN	Trie DAG + HMM
Speed	Slower (15 minutes)	Faster (two minutes)
Note	No specific dictionary	Precise mode No specific dictionary

Problem of Jieba

- Names

```
[['發言人', '林俐婉', '內部', '調動', '由江巍峰', '接任'],  
['內部', '稽核', '主管', '林', '志強', '內部', '調動', '由', '莊文清', '接任'],  
['會計', '主管', '藍俊雄', '內部', '調動', '由林鴻名', '接任'],  
['內部', '稽核', '主管', '游本', '詮', '內部', '調動', '由', '曾筱茜', '接任'],  
['財務經理', '洪廷宜', '內部', '調動', '由', '王婷', '渝', '接任'],  
['研發', '主管', '吳政峰', '內部', '調動', '由', '朱清立', '接任'],  
['總經理', '高', '進義', '離職', '由陳譽', '接任', '發言人', '高', '進義', '離職', '由陳譽', '接任'],  
['改派',
```

- Thousandths

```
'01',  
 '/',  
 '02',  
 設置',  
 1',  
  ',  
  ',  
 '200',  
  ',  
  ',  
 '000',  
 股給',  
 元富',  
 不限',  
 用途',
```


Problem of CKIP

- Inconsistency in date segmentation

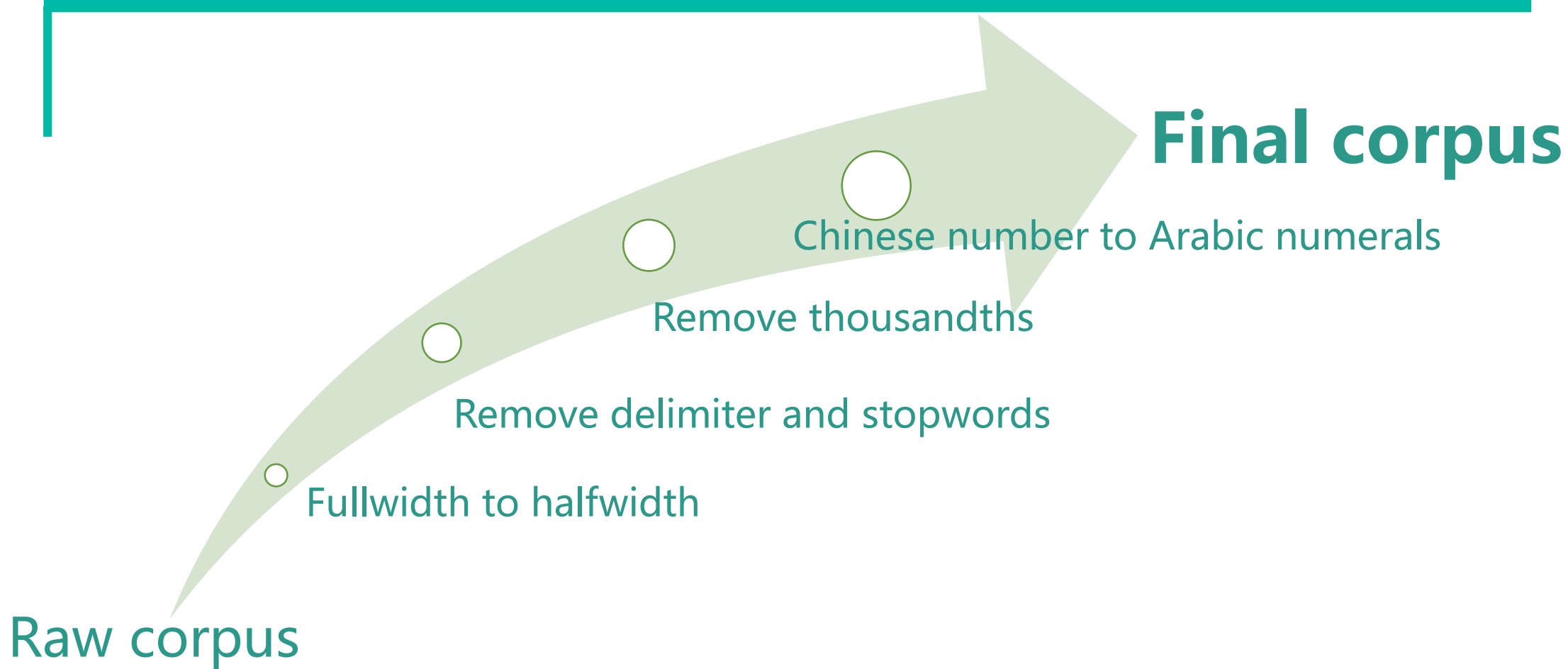
In [71]: words_list

```
'一般',  
'交易',  
'。',  
'。'],  
['內部', '稽核', '主管', '莊金維', '離職'],  
['2019',  
'/',  
'01/02',  
'收盤價',  
'17.00',  
'元',  
'',  
'月',
```

In [71]: words_list

```
'。'],  
['辭任', '1', '董', '。'],  
['董事長',  
'本人',  
'正邦',  
'投資',  
'2019/01/02',  
'設質',  
'1,200,000',  
'股',  
'給',  
'元富',
```

Other Data Preprocessing



Word Embedding – FastText

- Released by Facebook
 - An extension of the word2vec model
 - Understanding the meaning of “subword”
 - Based on characters instead of words
 - Transform each word from CKIP (Jieba) to embedding
 - E.g.: 發言人 林俐婉 內部 調動 由江巍峰 接任
- ➔ (6, 300) features



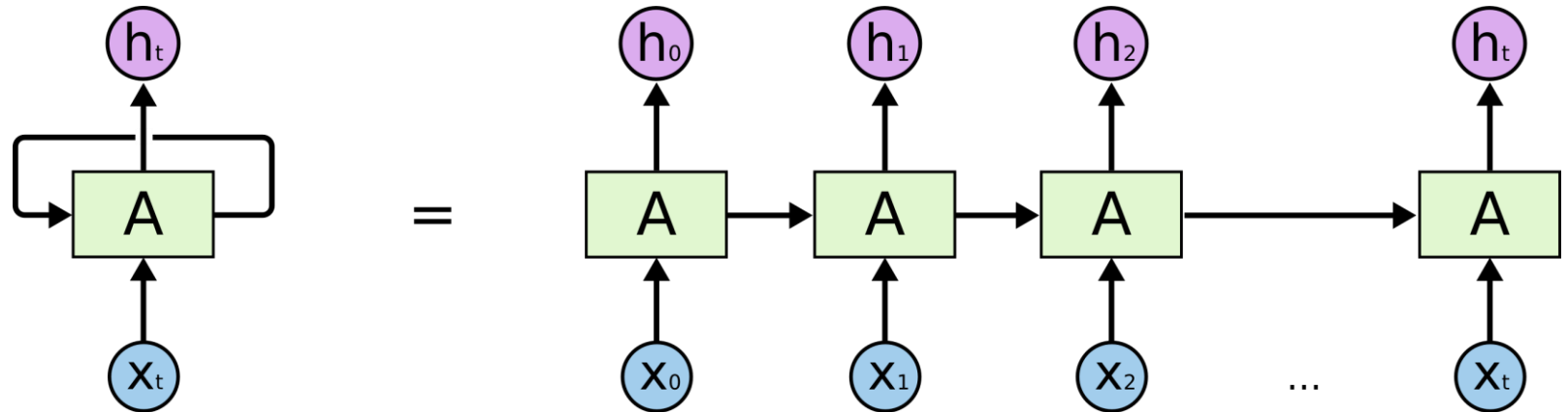
```
[[-0.149 , -0.1961,  0.1495, ...,  0.0319, -0.0883,  0.0485],  
 [-0.029 , -0.0061,  0.5102, ..., -0.1517, -0.0941, -0.0351],  
 [-0.1419,  0.0704,  0.4526, ..., -0.0045, -0.0592, -0.0627],  
 [ 0.2194, -0.0813,  0.3686, ..., -0.015 ,  0.0365, -0.0035],  
 [ 0.0073, -0.0087,  0.0978, ...,  0.1146,  0.0602,  0.0774],  
 [ 0.    ,  0.    ,  0.    , ...,  0.    ,  0.    ,  0.    ]]
```

Section III

Machine Learning Models

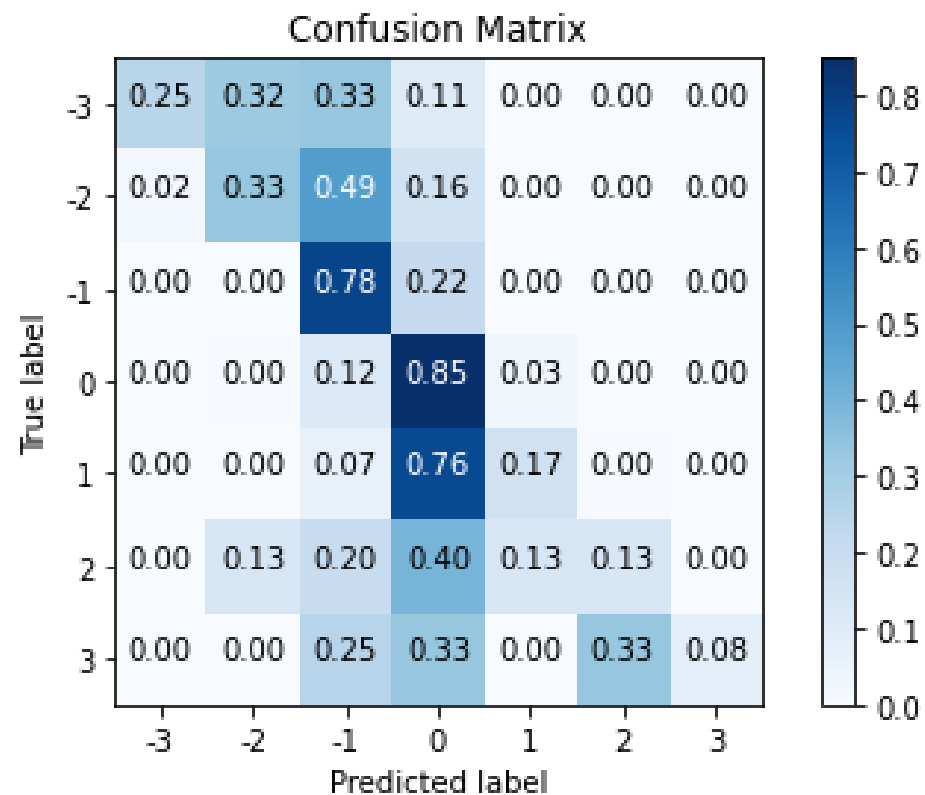
Recurrent Neural Network

- Extract information from a temporal sequence
- Applications
 - Stock prediction
 - Machine translation
 - Speech synthesis



Simulation Results – RNN

- Set max_sequence_length to 100
- Single layer LSTM
- Regression
- Train Acc: 0.8149423113296765
- Train MSE: 0.23257848426011588
- Test Acc: 0.6963507625272332
- Test MSE: 0.3689669900819255

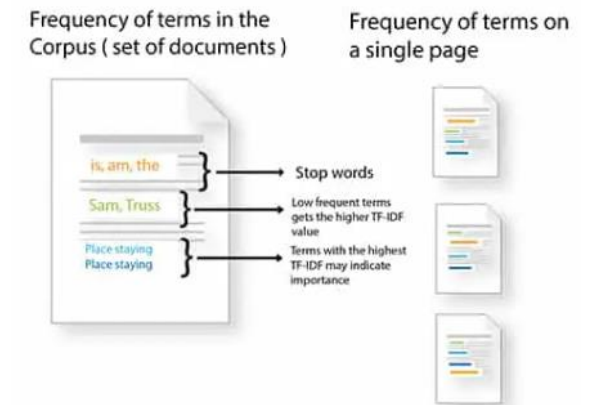


Problem of RNN

- A huge range of sequence length
 - max_sequence_length = 1442
 - min_sequence_length = 3
 - avg_sequence_length = 50
- RNN can't effectively extract information from the sentence
- The importance of words in each sentence is ignored

Term Frequency-Inverse Document Frequency (TF-IDF)

- Term frequency: $tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$
 - The frequency of t_i in document d_j
- Inverse document frequency: $idf_i = \log \frac{|D|}{|\{j:t_i \in d_j\}|}$
 - t_i is common or rare across all documents
- $tfidf_{i,j} = tf_{i,j} \times idf_i$
 - Evaluate the importance of t_i in document d_j
 - Filter out common terms and keep the important words



Sentence Embedding

- Use TF-IDF results to weight the word embedding in each sentence
 - TF-IDF: $|dataset| \times |bag_of_word| = (36717, 71590)$
 - Word embedding: $|bag_of_word| \times dimension = (71590, 300)$
- Sentence embedding: $TFIDF \times Word\ Embedding = (36717, 300)$
- Use 300 dimensions to represent each sentence
- E.g.: 發言人 林俐婉 內部 調動 由江巍峰 接任
➔ (1, 300) features

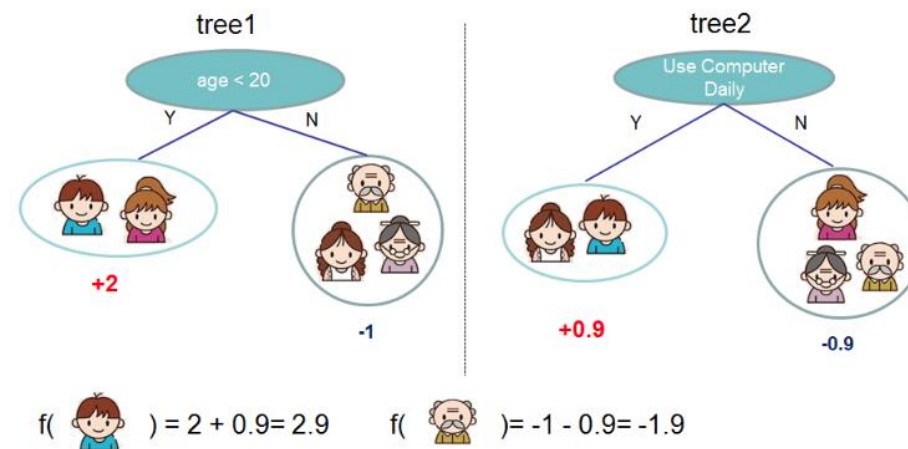
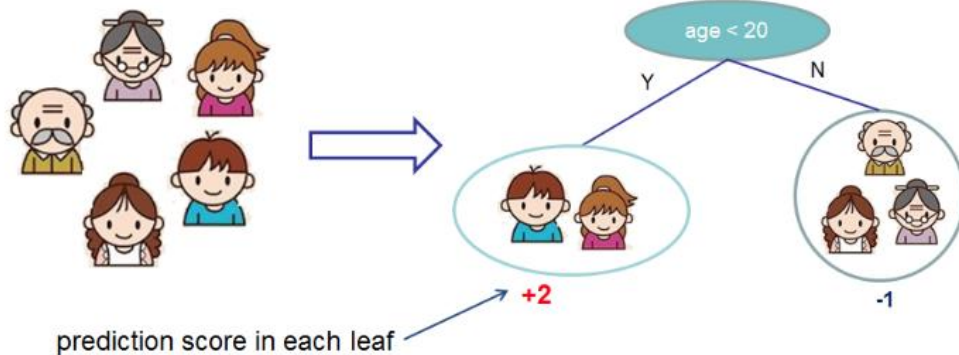
Extreme Gradient Boosting (XGBoost)

- Tree-based classifier
 - Ensemble of many weak prediction models

XGBoost
eXtreme Gradient Boosting

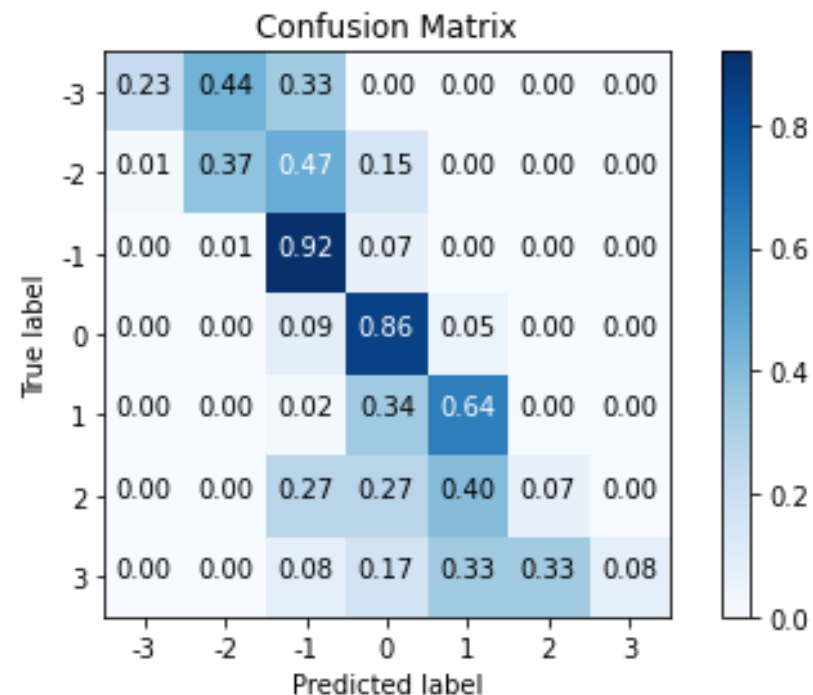
Input: age, gender, occupation, ...

Like the computer game X



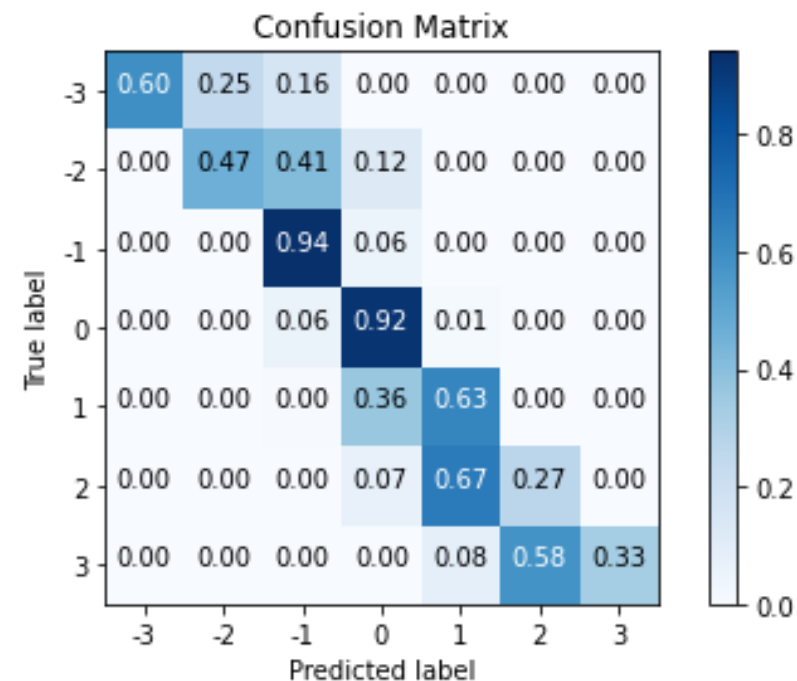
Simulation Results: XGBoost

- `xgb.XGBRegressor(max_depth=6, n_estimators=120)`
- Train Acc: 0.9738194940932149
- Train MSE: 0.030195149706289694
- Test Acc: 0.8404139433551199
- Test MSE: 0.17954414897565352
- Slight overfitting



Simulation Results: DNN

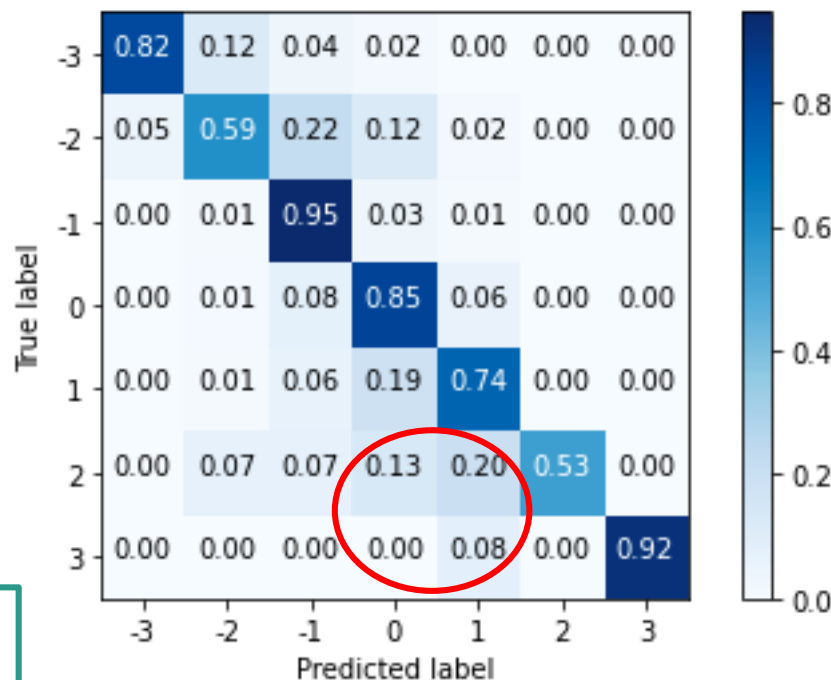
- Three hidden layers: (100, 50, 10)
- Regression
- Train Acc: 0.9078872706638925
- Train MSE: 0.08210409228625334
- Test Acc: 0.8748638344226579
- Test MSE: 0.12685892429858606
- Better than XGBoost



Simulation Results: K Nearest Neighbors (KNN)

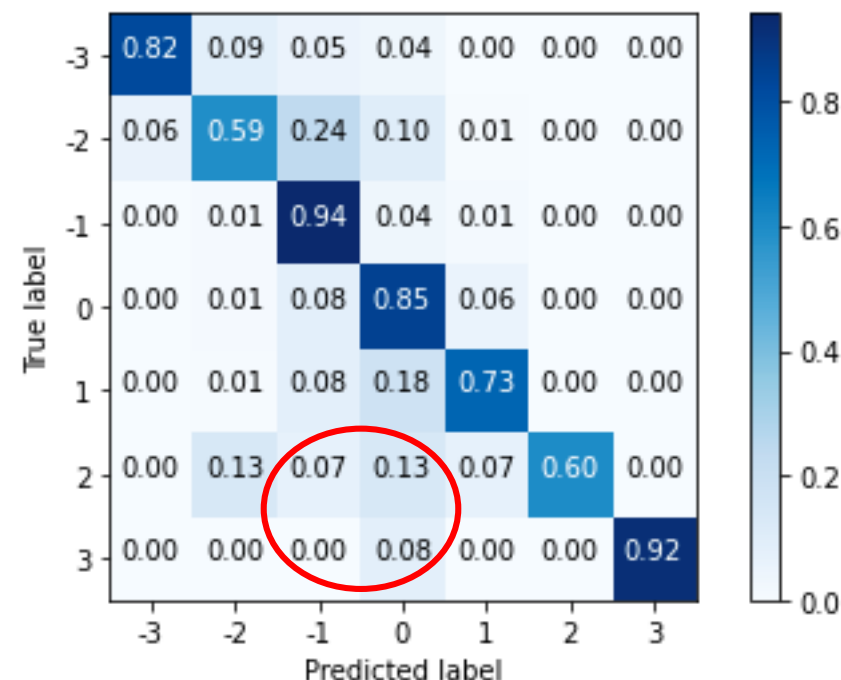
- K=3

CKIP Confusion Matrix



Test Acc: 0.8775871459694989
Test MSE: 0.21200980392156862

Jieba Confusion Matrix



Test Acc: 0.8759531590413944
Test MSE: 0.21541394335511982

Similar accuracy as NN,
but worse MSE

Problem of KNN

- “Imbalance data” → Hard to select the best “k” to enhance the accuracy of minority

2

緯創軟體第 2 季營收 14.09 億元，季增 9.64%、年增 47.57%，毛利率 26.06%，營益率 9.85%，稅後純益 1.43 億元，季增 73.63%、年增 64.47%，每股純益 2.37 元。上半年營收 26.93 億元，年增 52%，毛利率 25.04%，營。益率 8.82%，稅後純益 2.26 億元，年增 97.8%，每股純益 3.74 元。緯軟公布 7 月合併營收，達 4.77 億元，。月增 0.37 %、年增 40.95%，累計前 7 月合併營收 31.7 億元，年增 50.22%。緯軟上半年營收動能來自各地區的。業務成長包含中國、台灣、香港、日本及美國，上半年各地區營收占比為中國 54%、台灣及香港 28%、日本 15%、。其他地區則為 3%。

```
Out[41]: array([0.          , 0.33333333, 0.          , 0.          , 0.33333333,
                0.33333333, 0.          ])
```

```
Out[52]: array([0. , 0.2, 0.2, 0.2, 0.2, 0.2, 0. ])
```

- Hard to quantify Numerical Value

2019年10月累計營收833,569千元，年減18%。2019年10月單月營收124,687千元，年增68%。

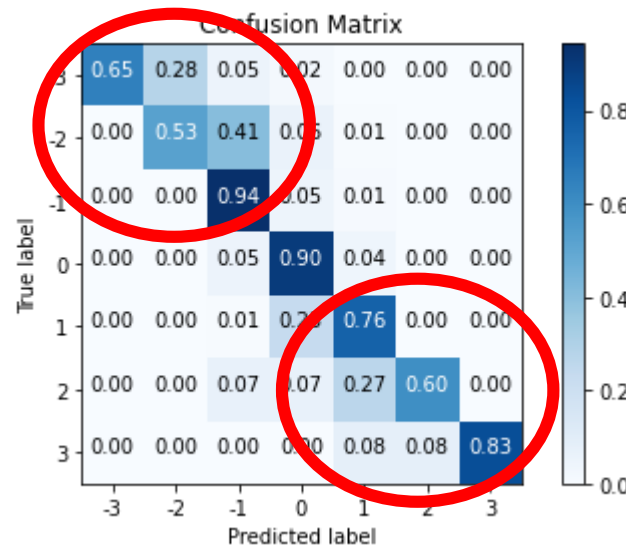
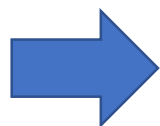
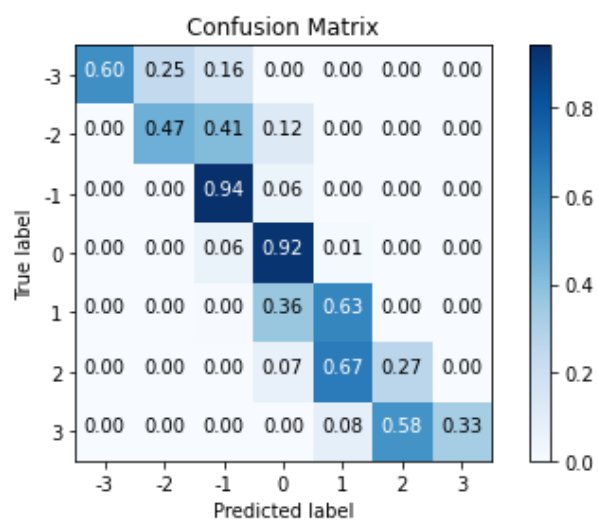
2019年12月累計營收626,923千元，年減2%。2019年12月單月營收65,838千元，年增61%。

Section IV

Feature Analysis

Feature Analysis: Big Event & Small Event

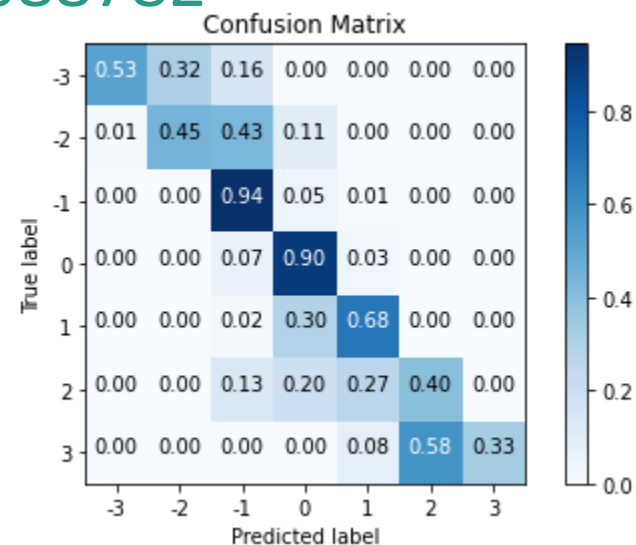
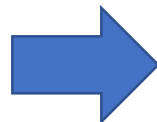
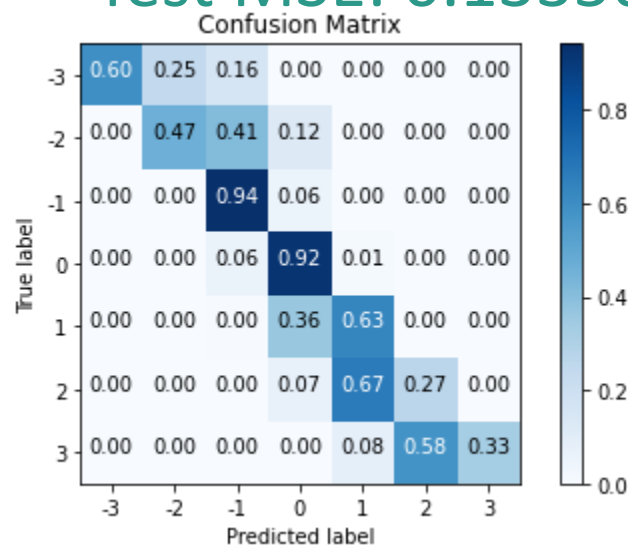
- Train Acc: 0.9309249101569889
- Train MSE: 0.06798529576694025
- Test Acc: 0.8924291938997821
- Test MSE: 0.11491382229155533



Effectively improve the prediction accuracy of important event

Feature Analysis: Stock

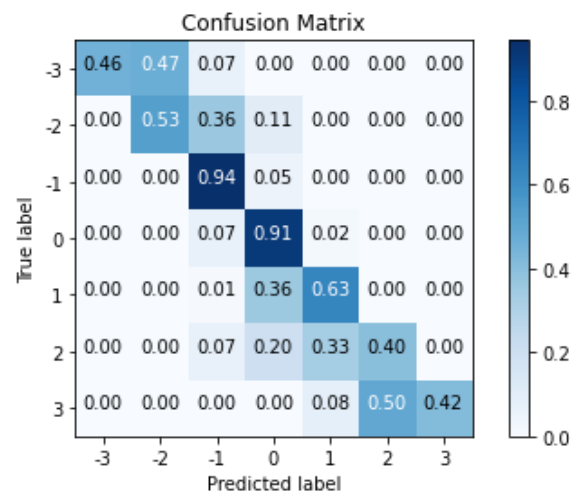
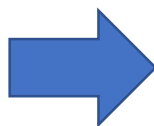
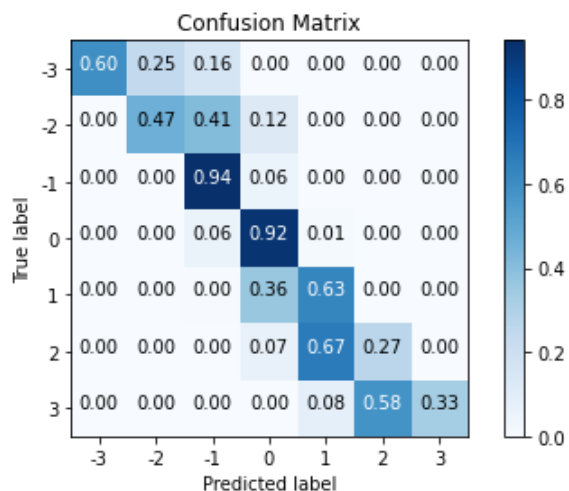
- Train Acc: 0.9264233024399471
- Train MSE: 0.06658710773337446
- Test Acc: 0.878131808278867 (0.8748638344226579)
- Test MSE: 0.1335647368088782



Only slight improvement

Feature Analysis: TCRI

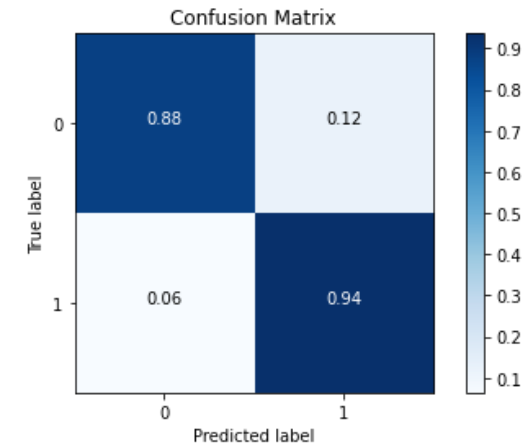
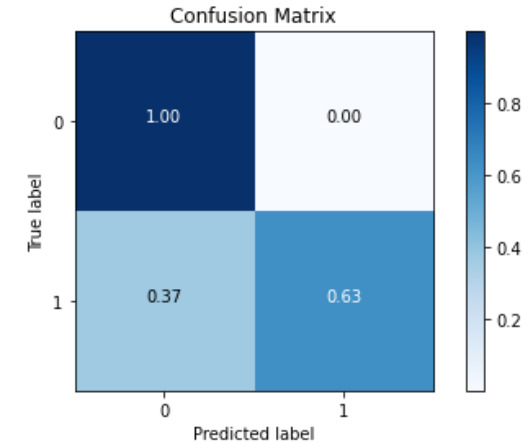
- Train Acc: 0.9082277283903916
- Train MSE: 0.08263161848890271
- Test Acc: 0.8737745098039216 (0.8748638344226579)
- Test MSE: 0.12831244750977194



Can't improve accuracy

Importance of Event Prediction

- Absolute value $\geq 2 \rightarrow$ important
 - Only 3.467% is important \rightarrow imbalance
 - Test Acc: 0.9848856209150327
 - Test MSE: 0.013307634
-
- Absolute value $\geq 1 \rightarrow$ important
 - 71.332% is important
 - Test Acc: 0.9195261437908496
 - Test MSE: 0.059115127





Section V

Conclusion & Future Work

Conclusion & Future Work

- Conclusion
 - Achieve about 87% prediction accuracy with small MSE
 - Preprocessing is important (KNN)
 - TF-IDF is useful
- Future Work
 - Dataset - imbalanced data
 - More powerful model - BERT

Work Assignment

- 鄧傑方 : TF-IDF, NN Modeling, PPT, Demo
- 連奕茹 : KNN, Video Making, PPT
- 林韋丞 : EDA, PPT
- 林聖硯 : Data Preprocessing, Video Making, PPT Designer

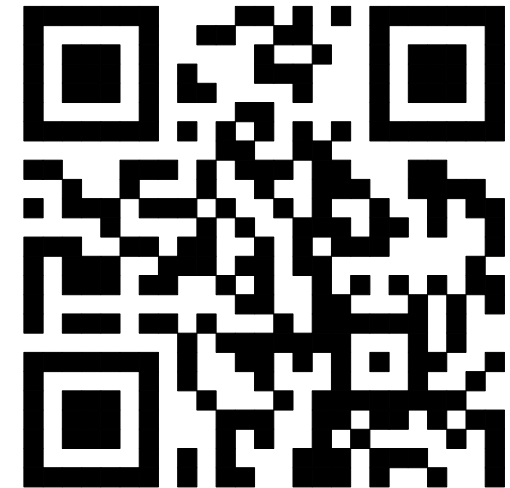
Other Information

Video



<https://youtu.be/G6nf6FLQOTA>

Demo



<http://140.112.20.131:1402/>

- Github link: <https://github.com/JieFangD/AI-News-Scoring-System>

Section VI

Appendix

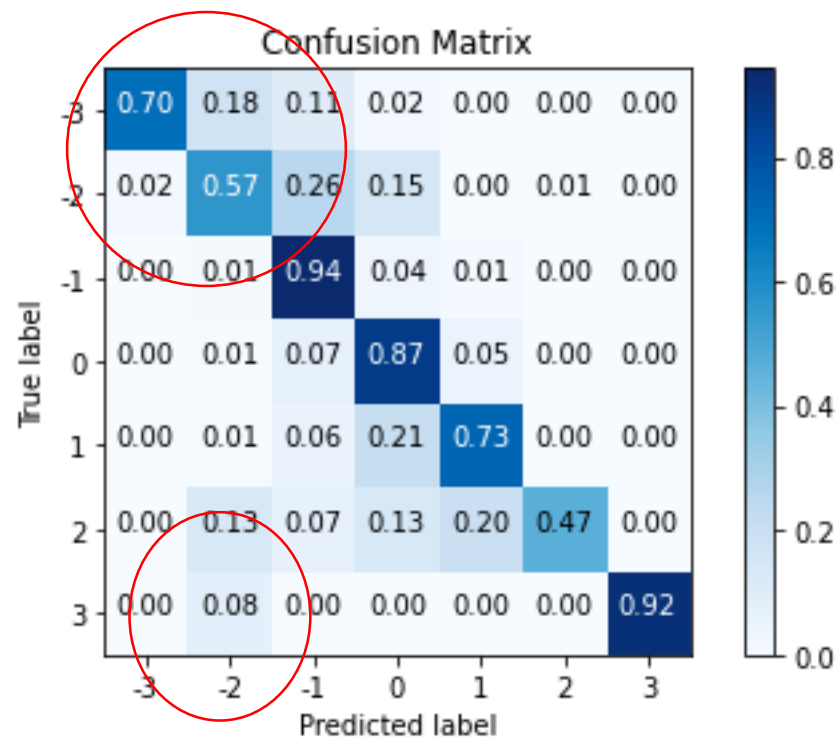


K Nearest Neighbors (KNN)

- Select the best “K”

```
'mean_test_score': array([0.85643278, 0.86565894, 0.8712423 , 0.86926769, 0.87114016,
    0.87001668, 0.86991455, 0.86865489, 0.86824635, 0.867225 ,
    0.86508018, 0.86371838, 0.86368434, 0.86187996, 0.86205018,
    0.86075648, 0.85983727, 0.85936064, 0.85878187, 0.85837334,
    0.85772648, 0.85684132, 0.85592211, 0.85571784, 0.85452627]),
'std_test_score': array([0.00211875, 0.00169534, 0.00312784, 0.00369274, 0.00347863,
    0.00331665, 0.00353533, 0.00370731, 0.00510545, 0.00384426,
    0.00379534, 0.00376731, 0.00392009, 0.00361048, 0.00341584,
    0.00459027, 0.00514016, 0.00517077, 0.00484017, 0.00532159,
    0.00563123, 0.0049465 , 0.00471049, 0.00503715, 0.00548359]),
'rank_test_score': array([22,  9,  1,  5,  2,  3,  4,  6,  7,  8, 10, 11, 12, 14, 13, 15, 1
6,
    17, 18, 19, 20, 21, 23, 24, 25], dtype=int32))
```

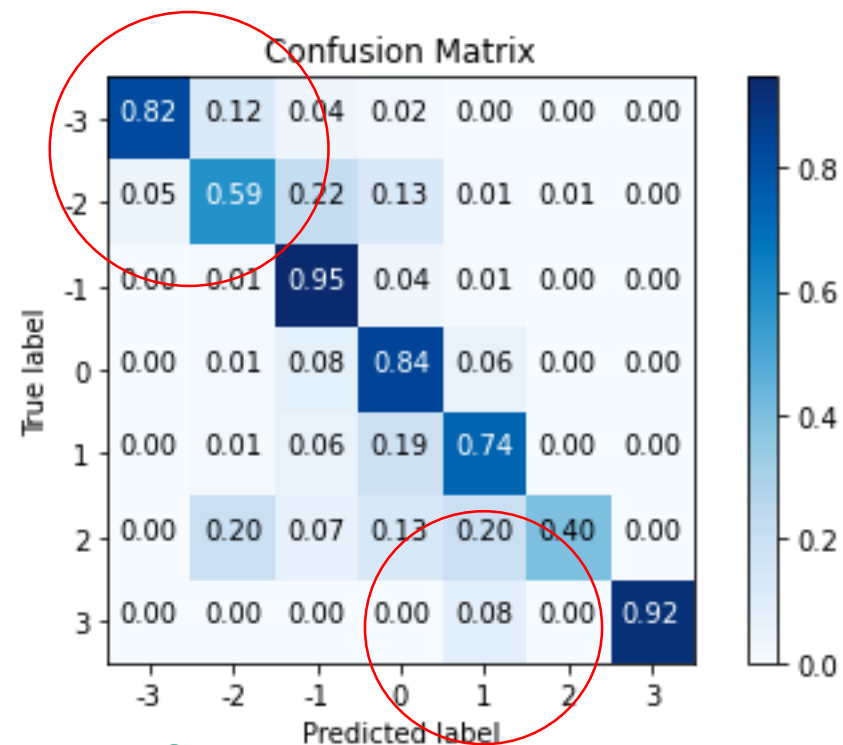
CKIP: K=3 vs K=5



K=5:

Test Acc: 0.880718954248366

Test MSE: 0.20179738562091504

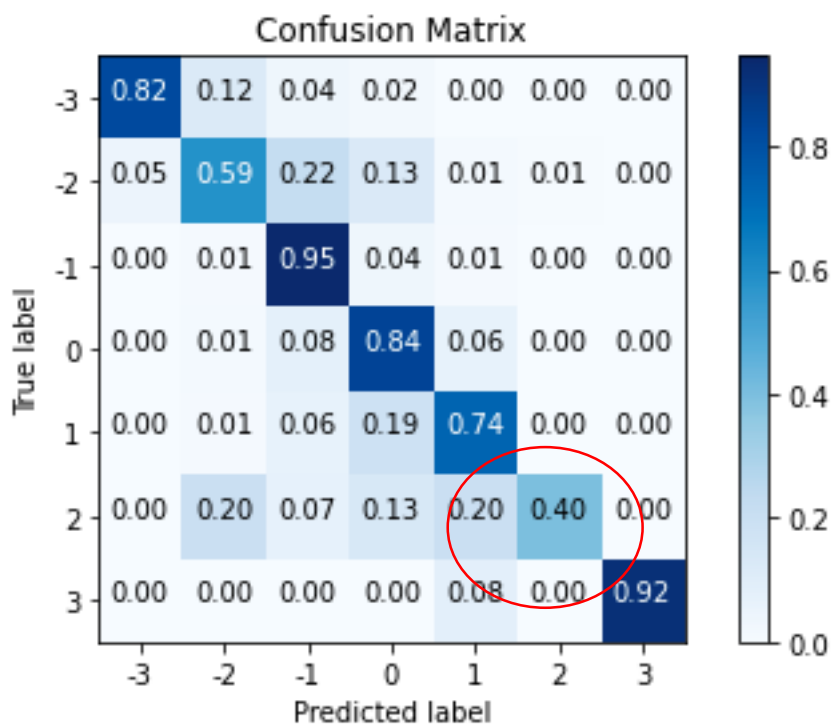


K=3:

Test Acc: 0.8766339869281046

Test MSE: 0.21840958605664487

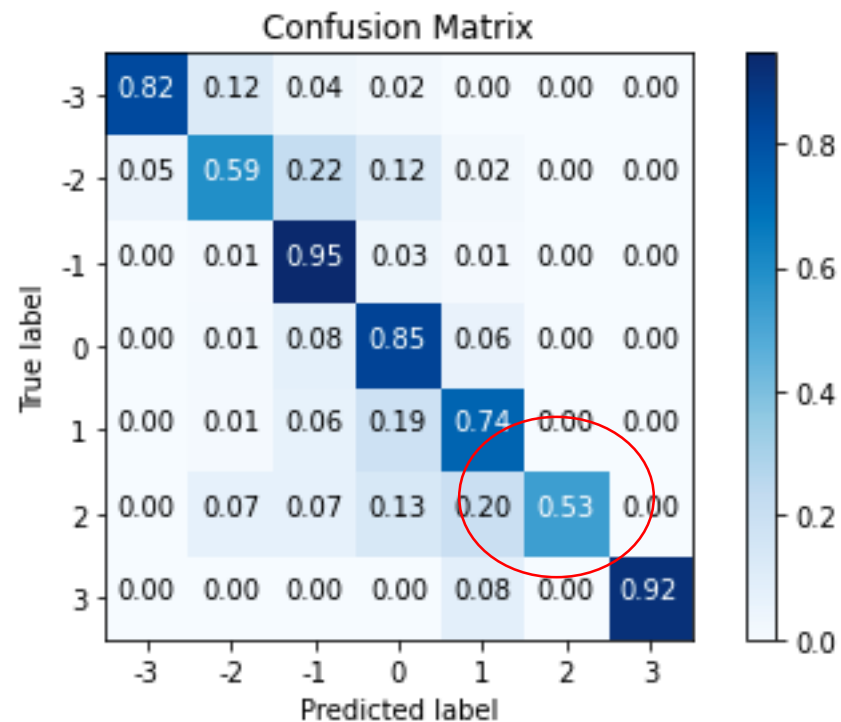
Improvement of CKIP



K=3:

Test Acc: 0.8766339869281046

Test MSE: 0.21840958605664487

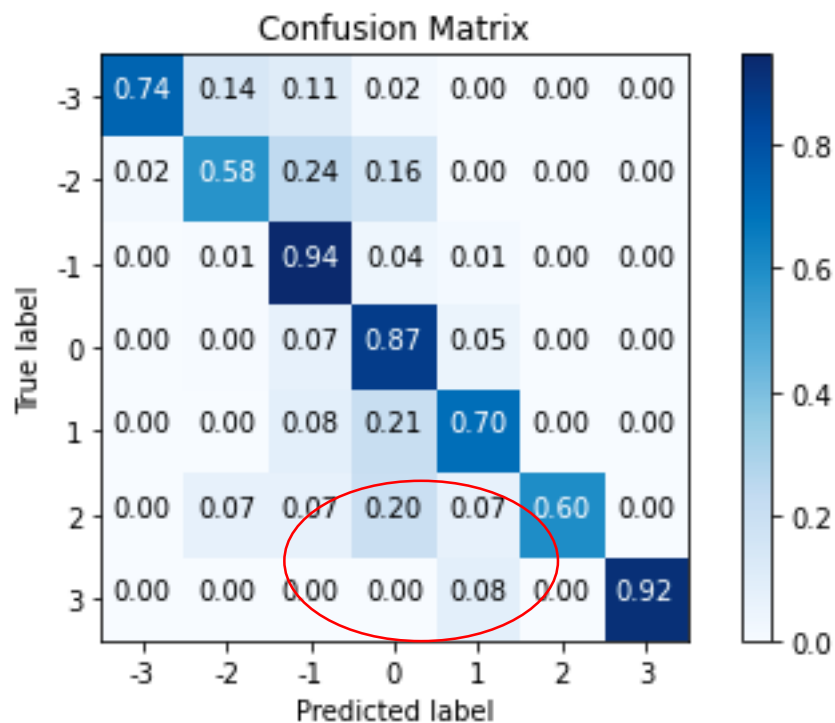


K=3:

Test Acc: 0.8775871459694989

Test MSE: 0.21200980392156862

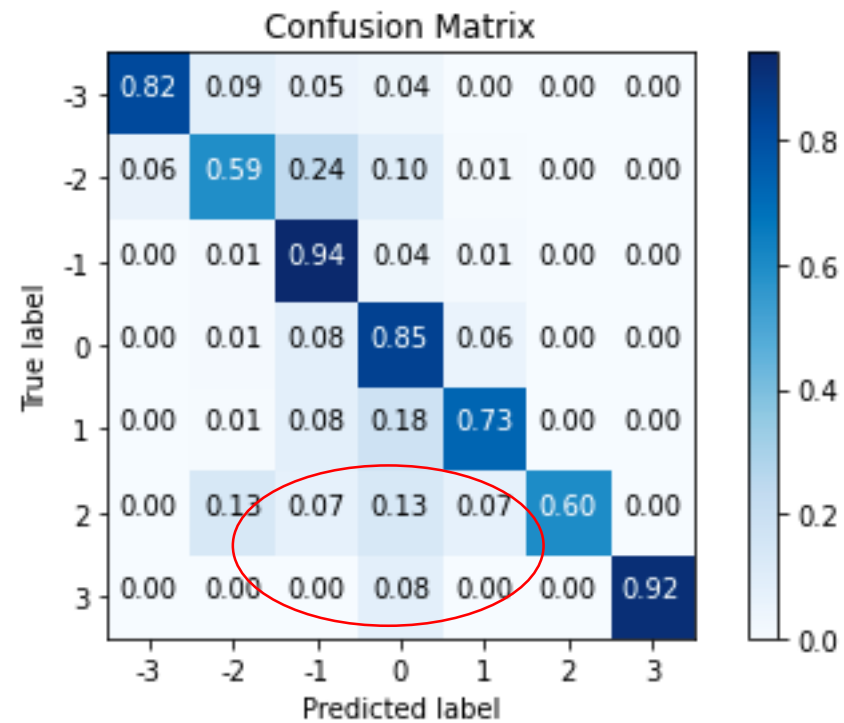
Jieba: K=3 vs K=5



K=5

Test Acc: 0.8779956427015251

Test MSE: 0.2079248366013072



K=3:

Test Acc: 0.8759531590413944

Test MSE: 0.21541394335511982