

悲傷 Peter 與他的快樂小伙伴

SMA Final Project

組員

B094020046 黃奕瑋

B094020011 邱亮傑

B094020007 陳文薇

M124020010 鄭雅云

Overview

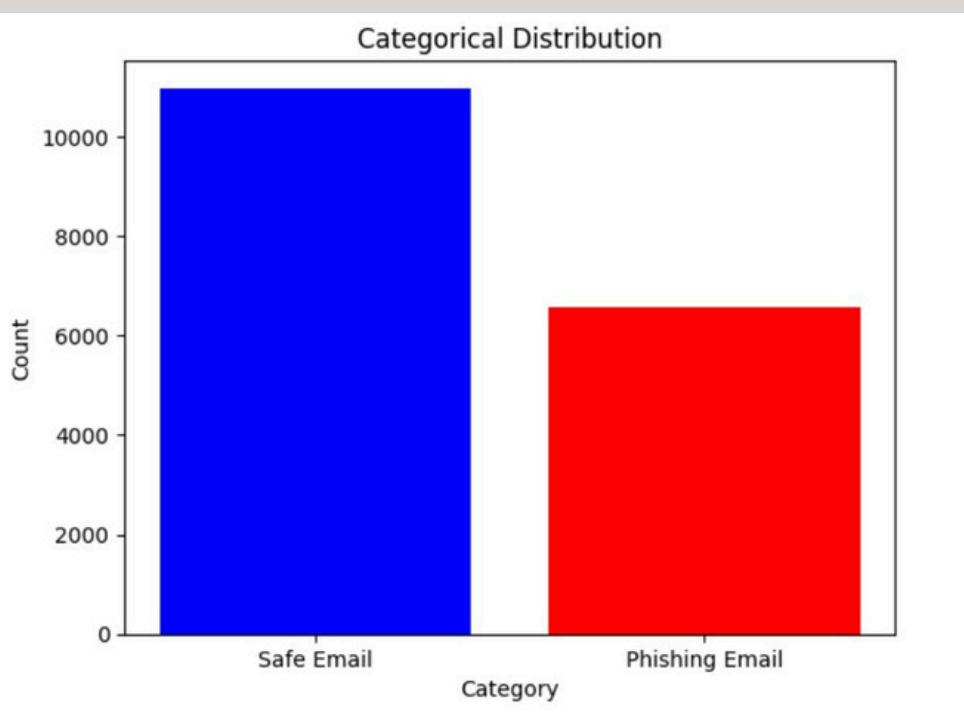
- 01 EDA、詞頻分析
- 02 詞性分析
- 03 主題模型
- 04 情緒分析
- 05 分類器
- 06 總結

EDA

了解資料集

資料集：Kaggle 上的釣魚信件資料集

- 資料欄位數：3 個欄位 – index, Email Text, Email Type
- Email Type: Safe Email, Phishing Email
- 資料筆數：18650 筆



資料來源：
[https://www.kaggle.com/datasets/
subhajournal/phishingemails/data](https://www.kaggle.com/datasets/subhajournal/phishingemails/data)

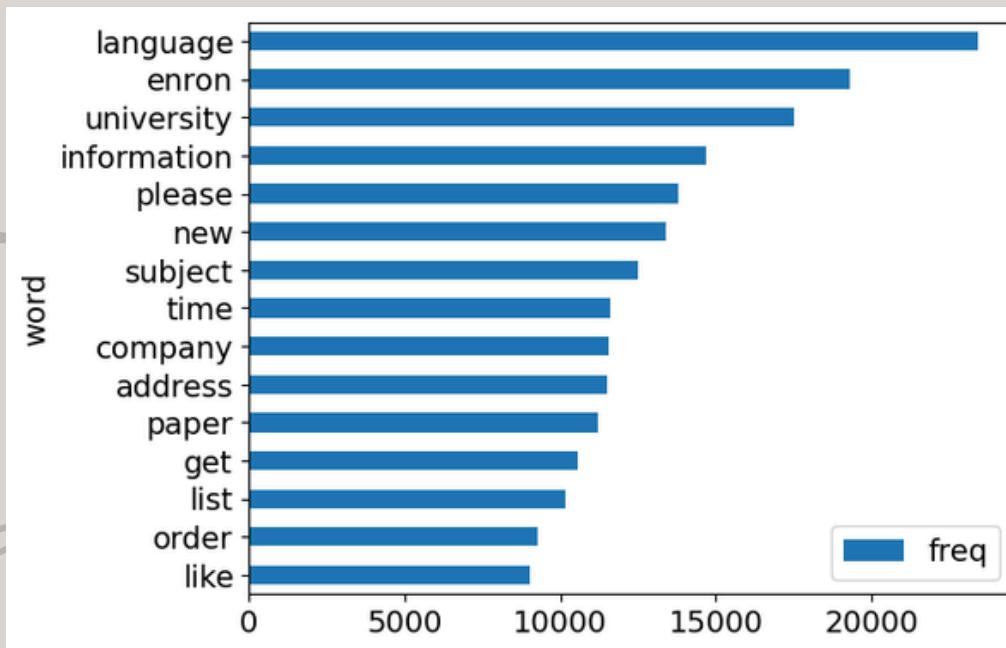
平均字數	535.17
字數標準差	25850.65
最少字數	0
25% 字數	74
50% 字數	160
75% 字數	354
最多字數	3527576

詞頻分析

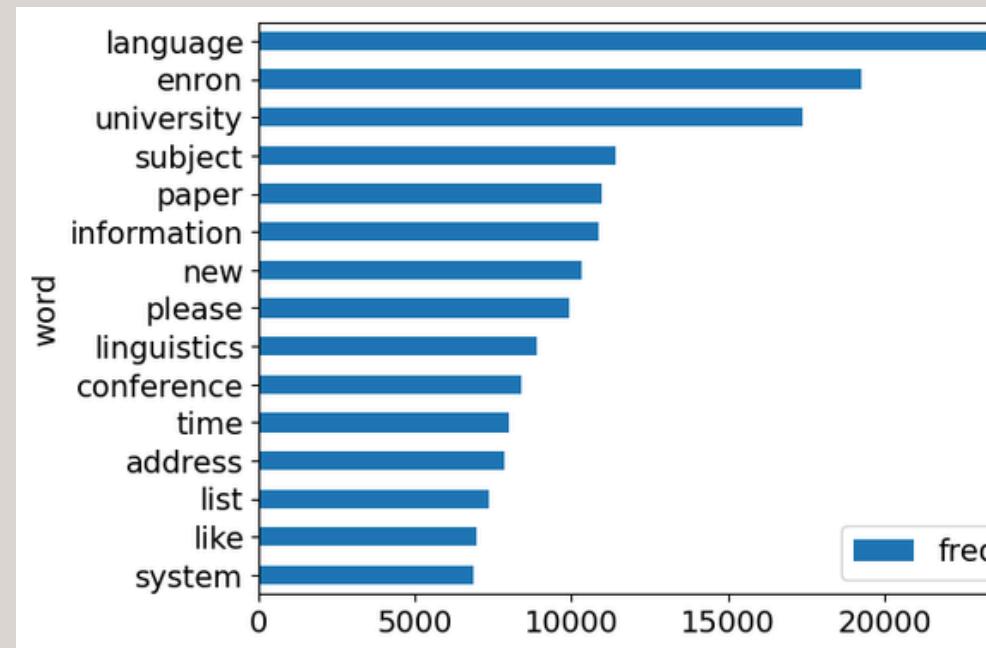
計算去除停用字後的斷詞詞頻

分類觀察：總信件、安全信件、釣魚信件

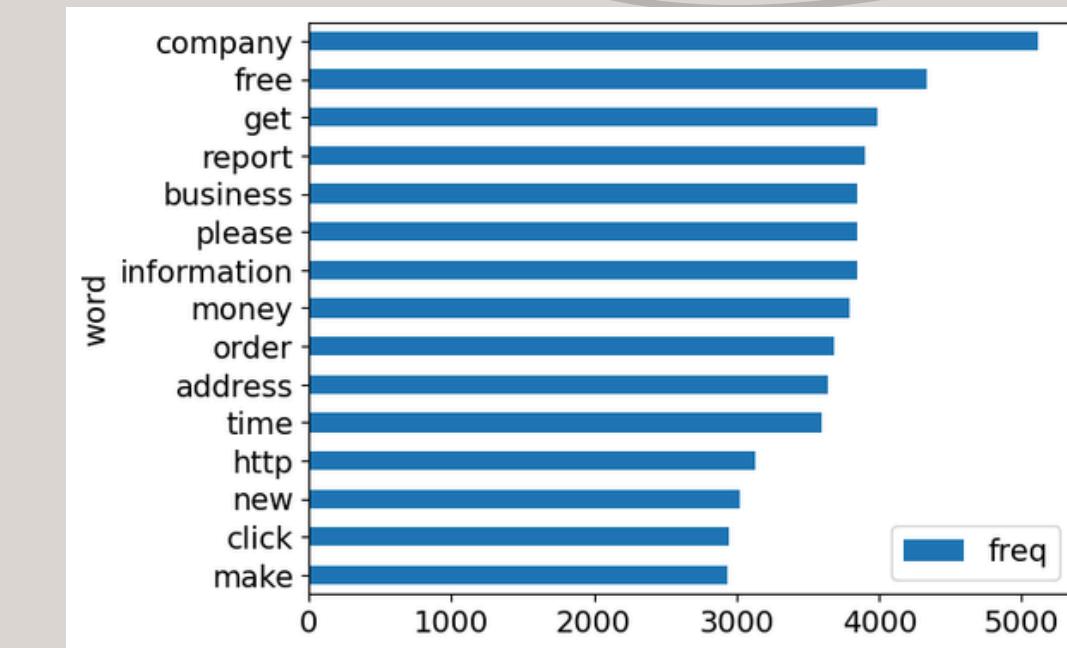
詞頻分析結果與後續分析結果幾乎一致：釣魚信件具明顯特徵



- enron 事件
- 使用者設定 language



- enron 事件
- 使用者設定 language



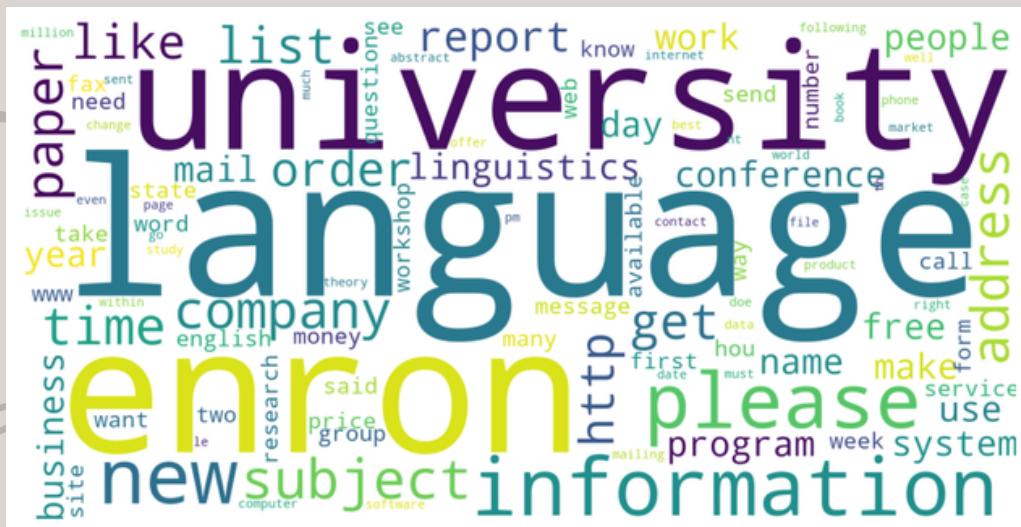
- company
- free、get、click、money

詞頻分析

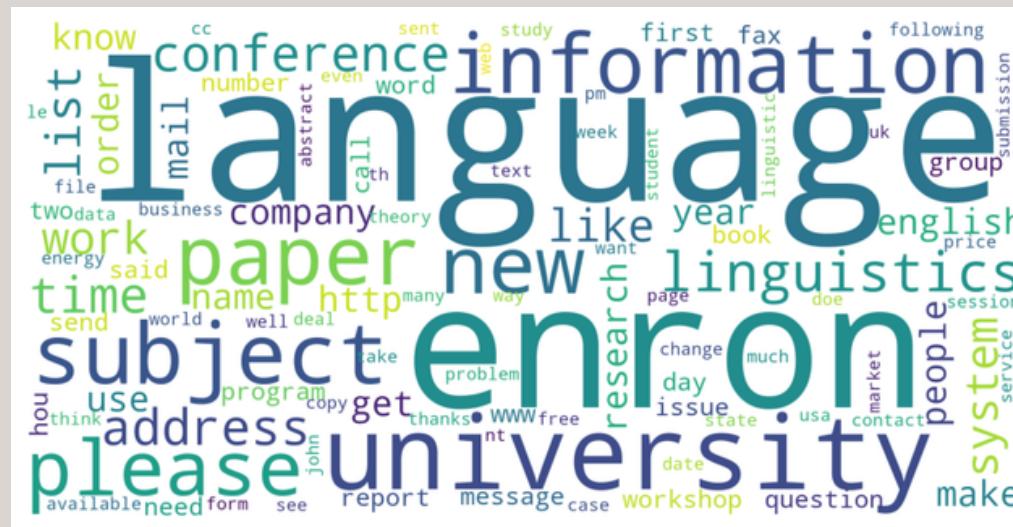
計算去除停用字後的斷詞詞頻

分類觀察：總信件、安全信件、釣魚信件

詞頻分析結果與後續分析結果幾乎一致：釣魚信件具明顯特徵



- enron 事件
 - 使用者設定 language



- enron 事件
 - 使用者設定 language

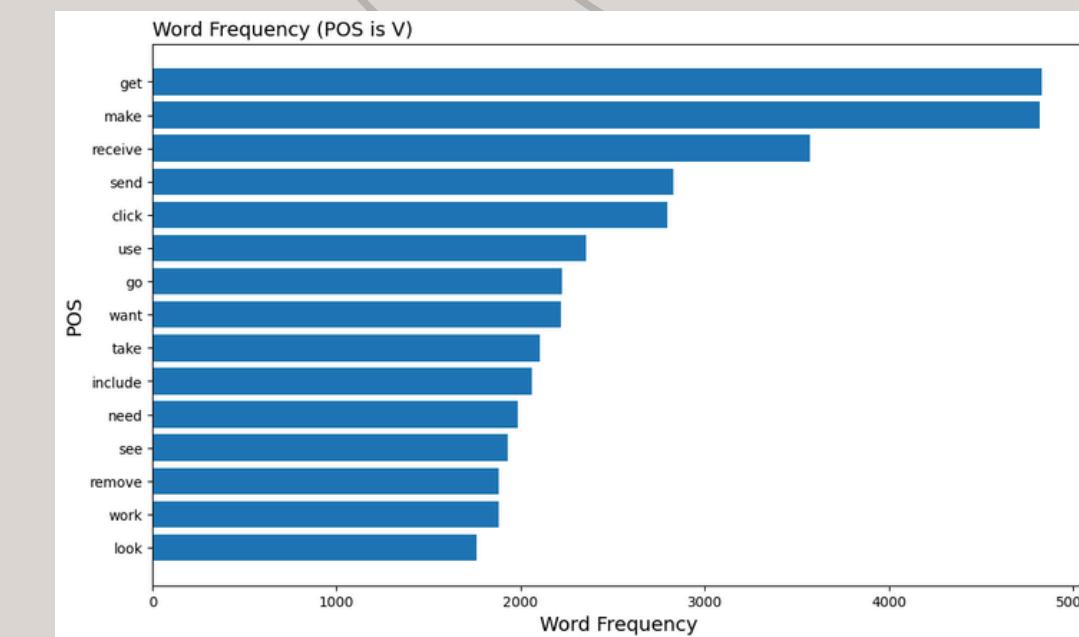
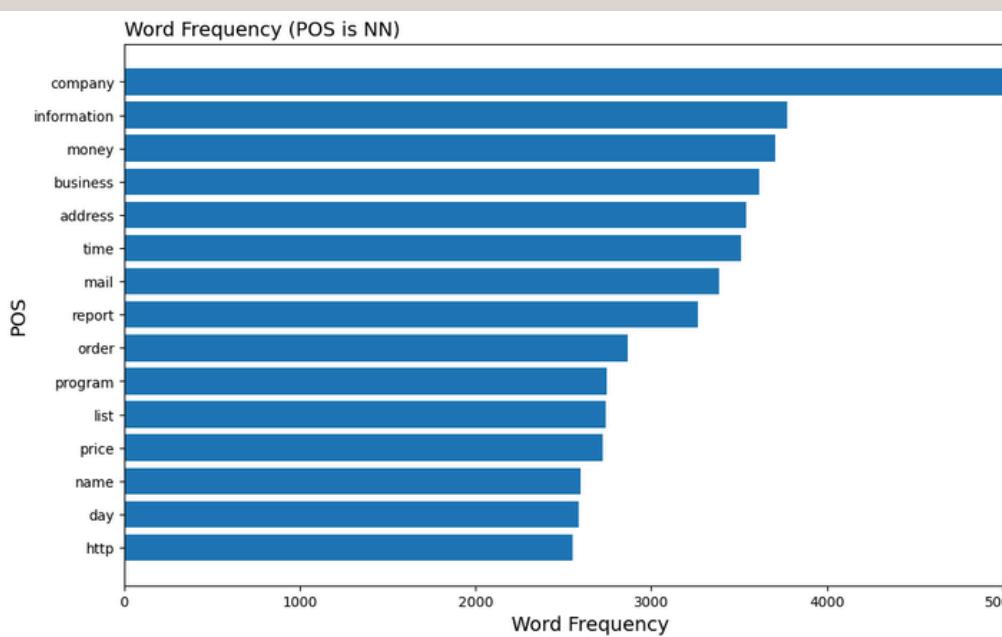
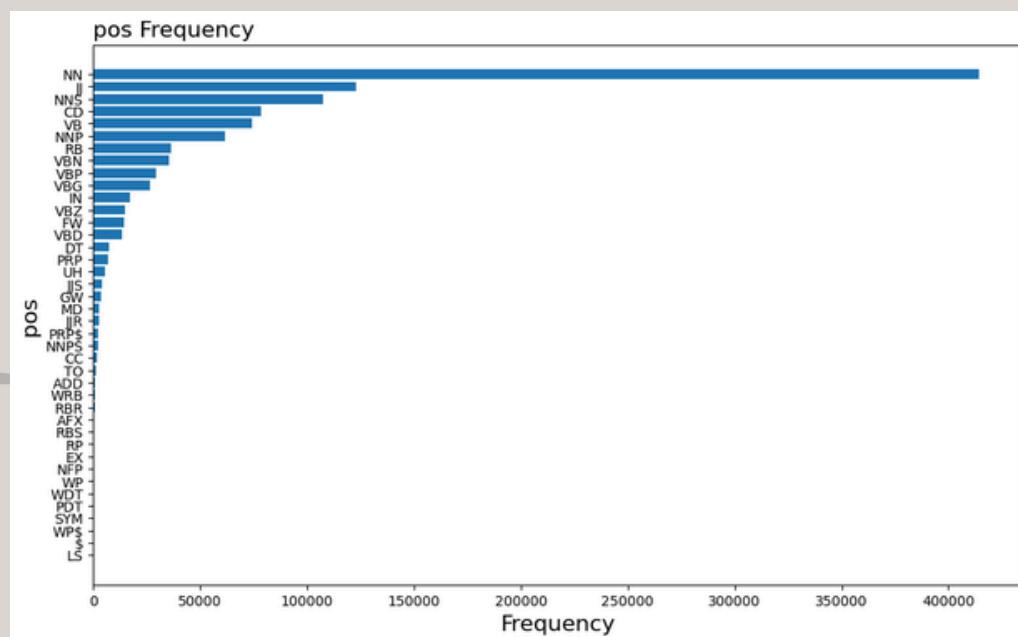


- company
 - free、get、click、money

POS 詞性分析

分析 lemmatized 後的釣魚信件內容的POS詞性

分類觀察：總釣魚信件 POS 分布、各詞性相關字詞



- 合併 NN 與 NNS = NN
- 合併 VB, VBN,... =V
- 不會納入 CD 基數

- company, business
- information, money, time

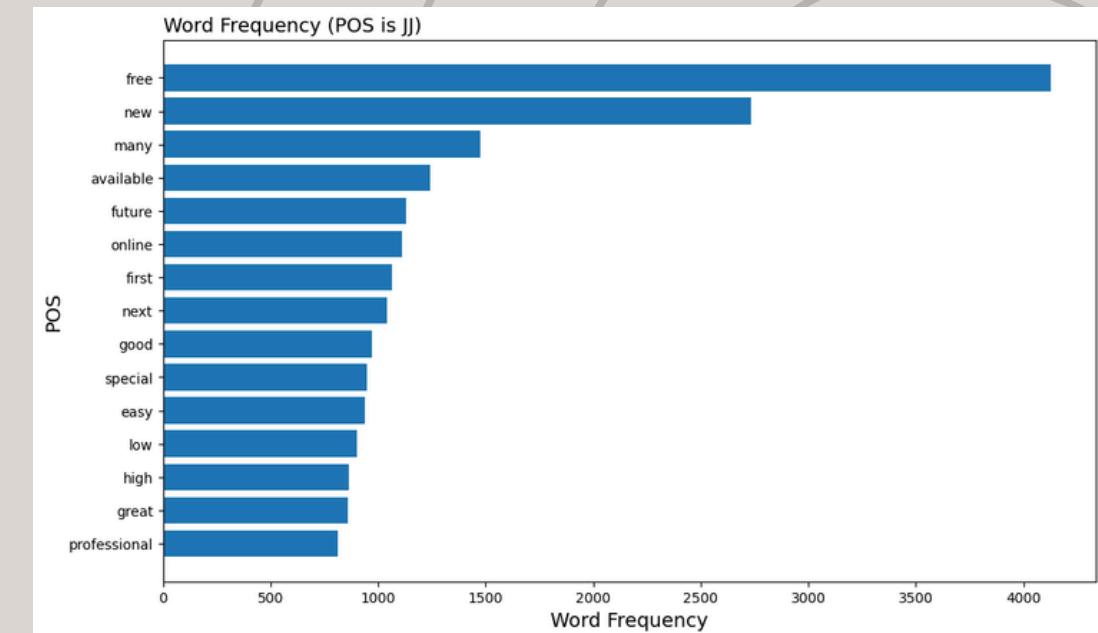
- get, make
- receive, send, click

POS 詞性分析

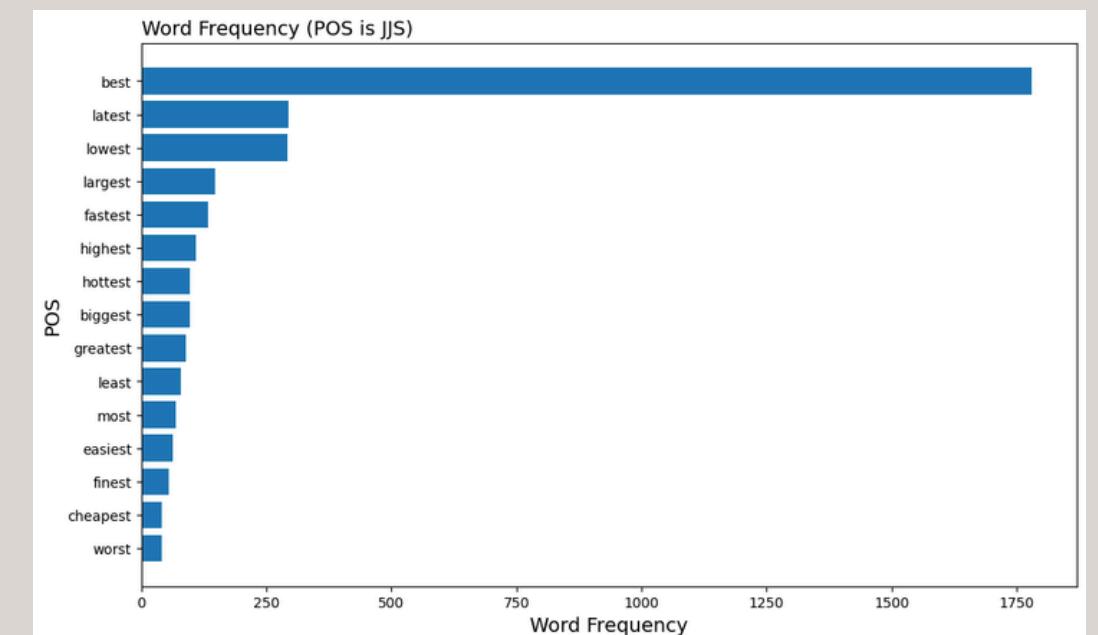
分析 lemmatized 後的釣魚信件內容的POS詞性

分類觀察：總釣魚信件 POS 分布、各詞性相關字詞

1. 提到公司 (company、business) 並給予收信者金錢 (money) 上的優惠，例如免費折扣 (free)
2. 強調這是最好 (best)、最大 (largest) 的機會或最低 (lowest) 的折扣，少部分則強調最新的優惠 (latest)
3. 請收信人點擊連結 (click)、收下優惠 (receive)、寄出回信 (send) 等，以獲得更多的資訊 (information)



- free
- new, many, available

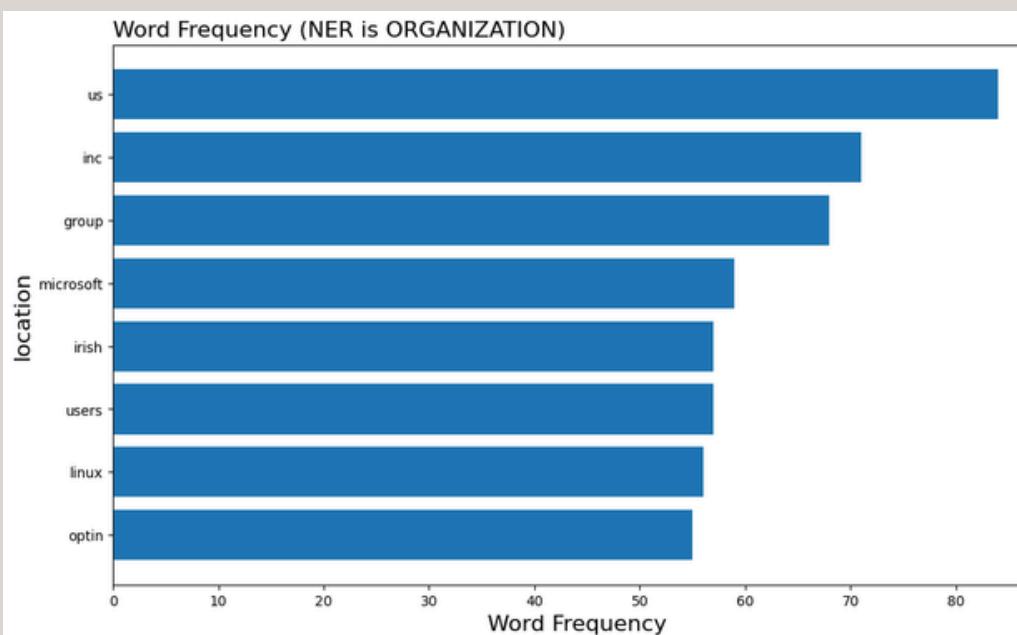
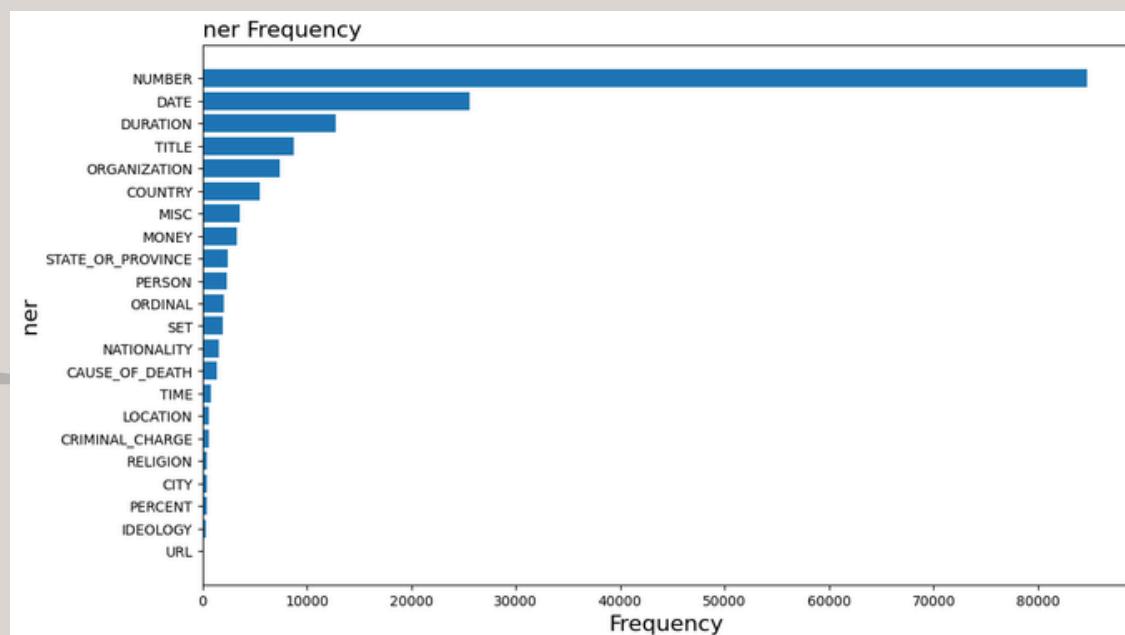


- best
- latest, lowest, largest

NER 詞性分析

分析 lemmatized 後的釣魚信件內容的NER詞性

分類觀察：總釣魚信件 NER 分布、各詞性相關字詞



index	Email Text	Email Type
0	744 intel play me 2 web cam 13 90 play me 2 web...	Phishing Email
1	858 lotto information office of the vice presi...	Phishing Email
2	1062 we have 800 expensive softwares for u to choos...	Phishing Email
3	1231 name get the best softwares for only 15 dkvn...	Phishing Email
4	1993 paydshl get latest softwares 99 savings ayo...	Phishing Email
5	2093 re your needed sofftwares at rock bottom pr...	Phishing Email
6	2315 fwd your needed sofftwares at rock bottom pr...	Phishing Email
7	2450 re we might have just what you need special ...	Phishing Email
8	2777 To avfsfazekashu Attn Marketing Department ...	Phishing Email
9	3130 award winning notification forehigh lottery in...	Phishing Email

- NUMBER: money, time
- DATE: latest

- 美國
- 知名軟體公司
- POS: company, business

- keyword: microsoft
- 1. 軟體、費用
- 2. 與詞頻分析和POS結果一致

LDA 主題模型

依據 perplexity 以及 pmi 找出最佳主題數

Perplexity 最低值：主題數 = 6

- 主題一、藥草與軟體
dagga、botanical、herb、adobe、corel
- 主題二、送錢與中獎
lottery、million、free、goldrush
- 主題三、金融投資
stock、investment、security、shareholder
- 主題四、藥品與網頁程式碼
cialis、viagra、pill、http、href、font
- 主題五、預付費詐騙有關主題
nigeria、lagos、deceased、foreigner、kin
- 主題六、減肥與亂碼
dieting、lean、fat、wor、kno

PMI 最高值：主題數 = 4

- 主題一、軟體
adobe、corel、macromedia、free、grant、get
- 主題二、藥草與保健食品
dagga、botanical、herb、dieting、aphrodisia
- 主題三、金融投資
company、stock、investment、investor
- 主題四、醫療藥物與網頁程式碼
cialis、xanax、dysfunction、http、href、font

GuidedLDA 主題模型

根據 LDA 分類結果定義 seed words

Seed Words

- 主題一、免費軟體
adobe、corel、macromedia、premiere、norton、acrobat、grant
- 主題二、金融投資
stock、investment、security、shareholder、investor、estimate、trading
- 主題三、醫療藥品
cialis、xanax、viagra、prozac、dagga、phentermine、aphrodisia
- 主題四、送錢與中獎
nigeria、deceased、foreigner、lottery、goldrush、moneymaking、nigerian
- 主題五、網頁程式碼
http、href、font、www、br、td、php

GuidedLDA 主題模型

GuidedLDA 分類結果

各主題代表字

- 主題一、免費軟體
free、address、order、report、get、business、
money、program
- 主題二、金融投資
company、statement、stock、security、
information、investment、market、inc
- 主題三、醫療藥品
get、click、free、offer、pill、http、online、life
- 主題四、送錢與中獎
account、money、fund、bank、number、claim、mr、
please
- 主題五、網頁程式碼
http、www、font、message、3d、price、color、info

GuideLDA 更有效地分離出我們關注的主題，我們也發現 free、get、click 等詞彙時常出現在各個主題中，這些詞彙的目的是利用優惠和金錢獎勵來吸引收信者點擊連結。

BERTopic 主題模型

BERTopic 分類結果

各主題代表字

- 主題一、釣魚信件的常見用詞
free、company、information、report、get
- 主題二、國家與政府
enenkio、islands、kingdom、marshall、atoll
- 主題三、科技技術
fuel、battery、cell、box、ones
- 主題四、網頁程式碼
function、documentwritett、easy、var、pattern
- 主題五、宗教
cns、counseling、christian、crditos、theological
- 主題六、學術領域
karpenkov、dcenter、align、occurrences、conductors
- 主題七、健康與醫學
acts、organs、pathway、endocrine、neuro

BERTopic 更詳細的分割不同主題，我們可以知道除了 LDA 所分出的常見主題外，釣魚信還包含了各式各樣的主題。這也提醒我們需要更加警覺，因為釣魚信會變換多種形式和內容進行釣魚。

主題模型結論

LDA、GuideLDA、BERTopic

透過主題分析，我們得出以下推論：

- 釣魚信常常出現如 company、free、get、click、money 等，目的是透過優惠和金錢來引誘收信者點擊連結。
- 釣魚信通常包含網址連結或網頁程式碼。
- 釣魚信大致分為免費軟體、金融投資、醫療藥品及給予金錢這四種常見類型。

通過主題分析，我們得以更深入了解釣魚信的特徵，從而更有效地辨識釣魚信件。

Sentiment Analysis

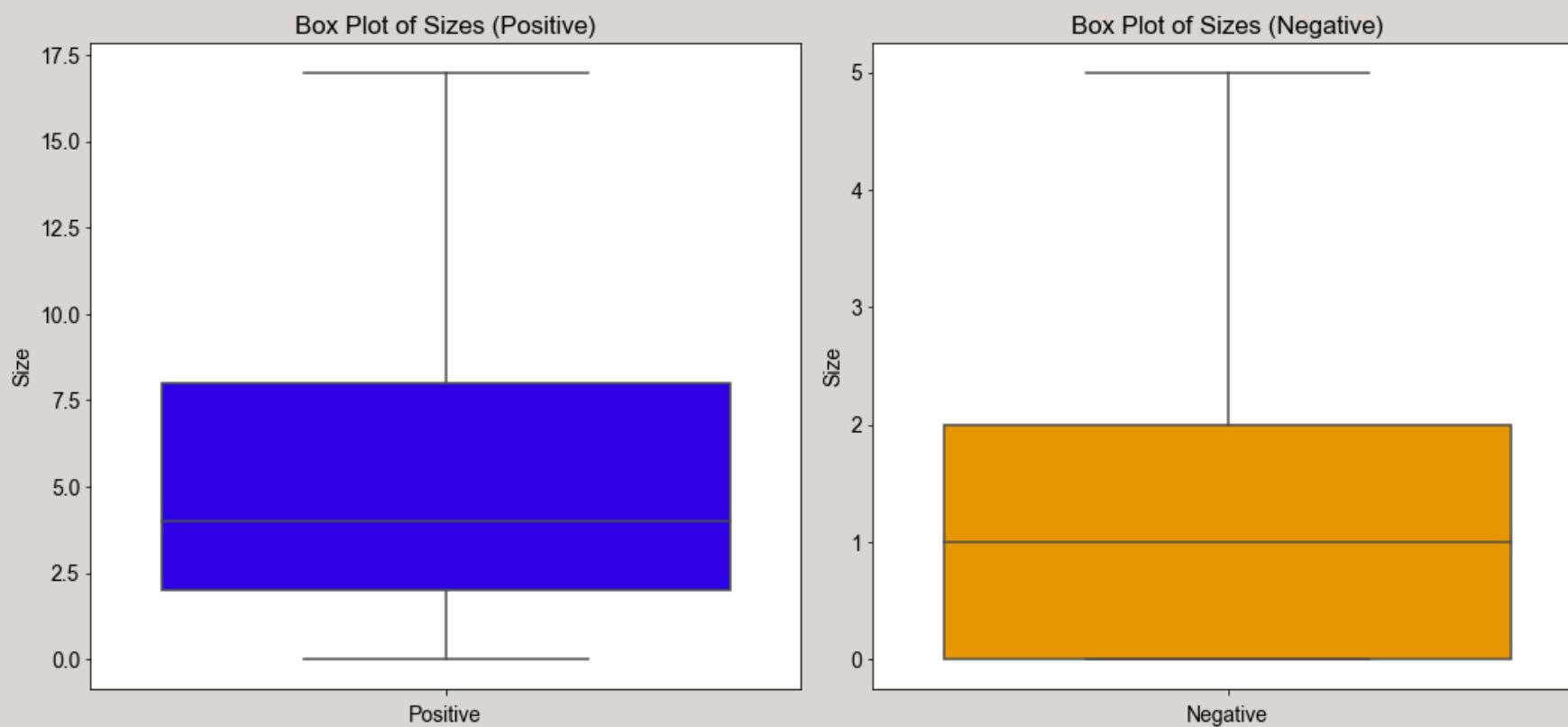
01 Lexicon

02 Bert

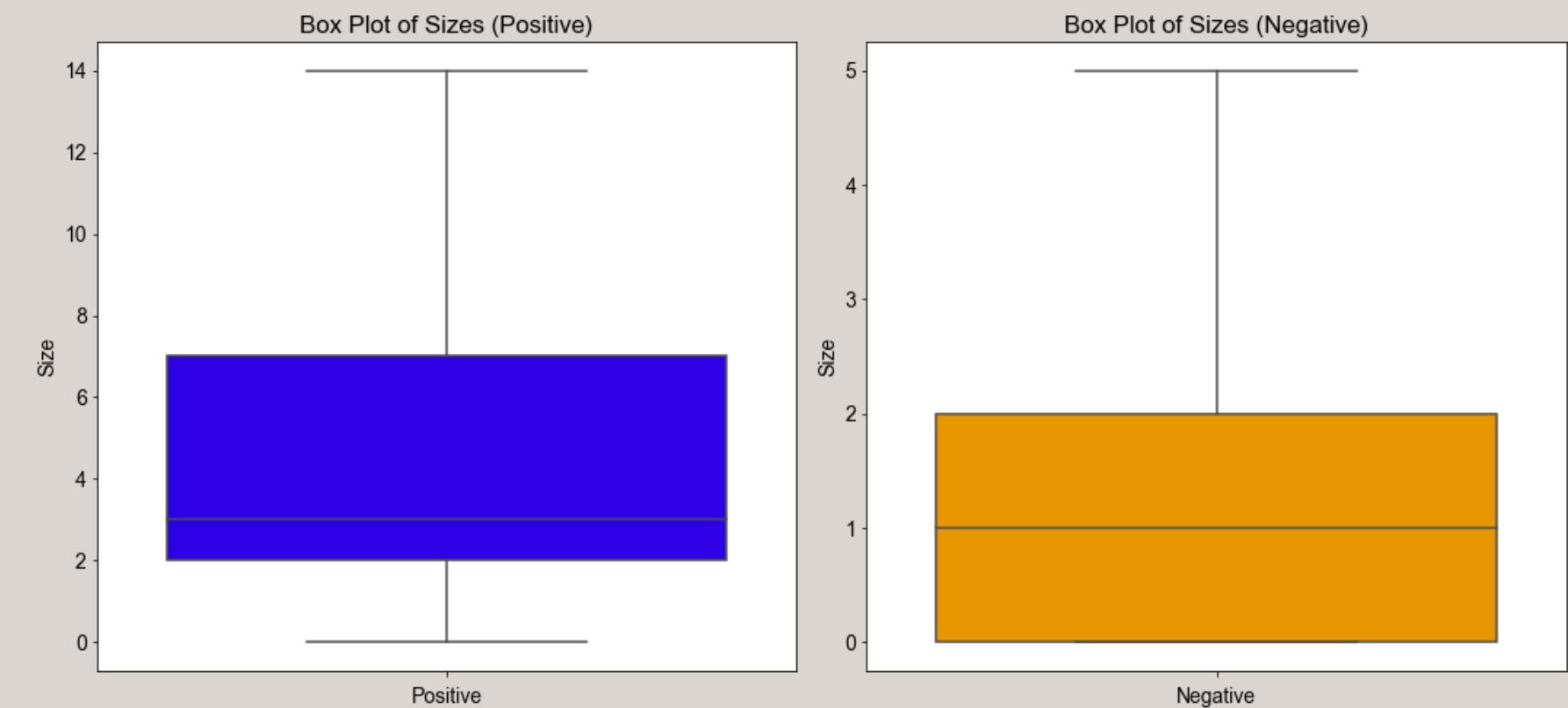
03 Gemma(LLM)

箱型圖

➤➤➤ 釣魚信件



➤➤➤ 安全信件

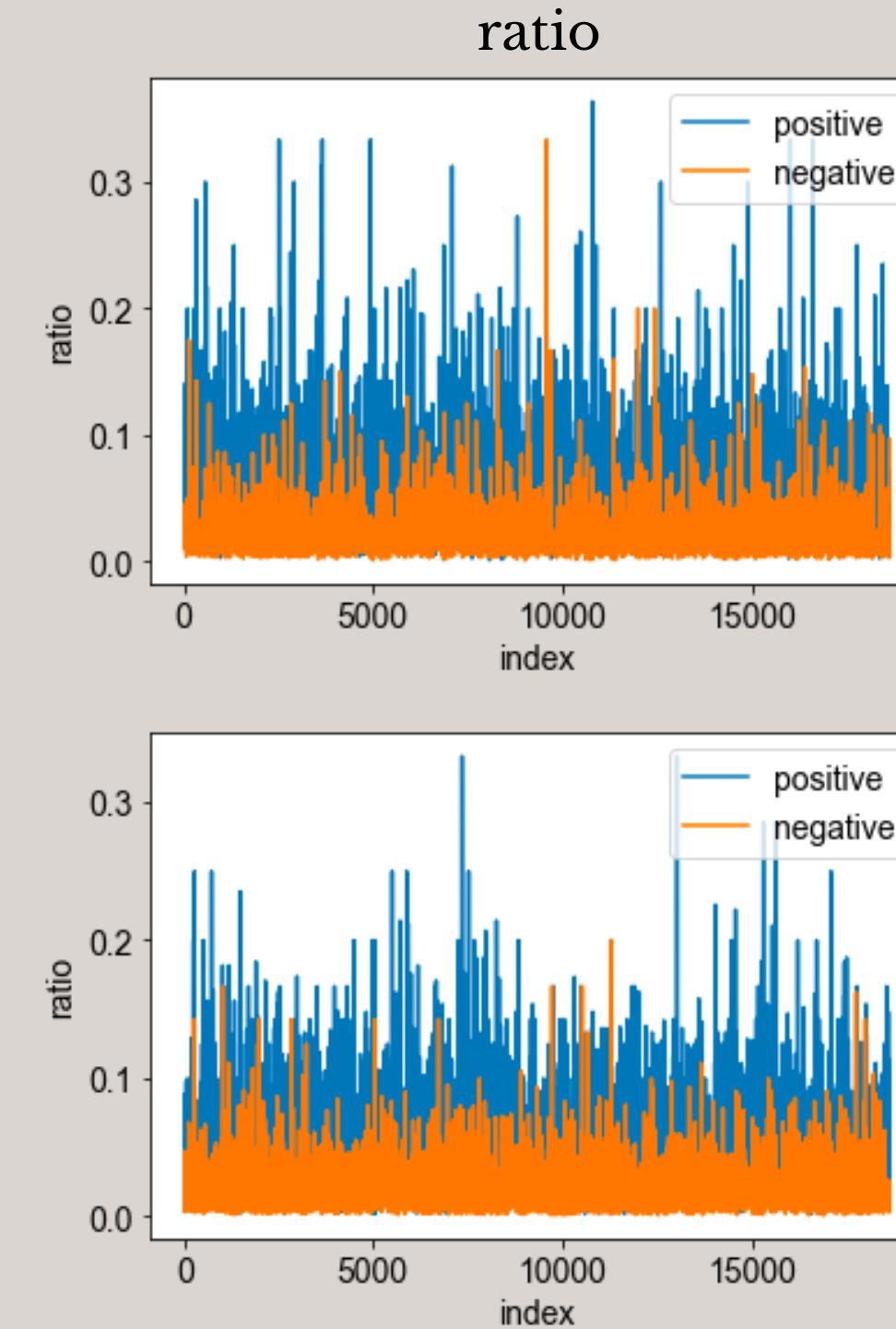
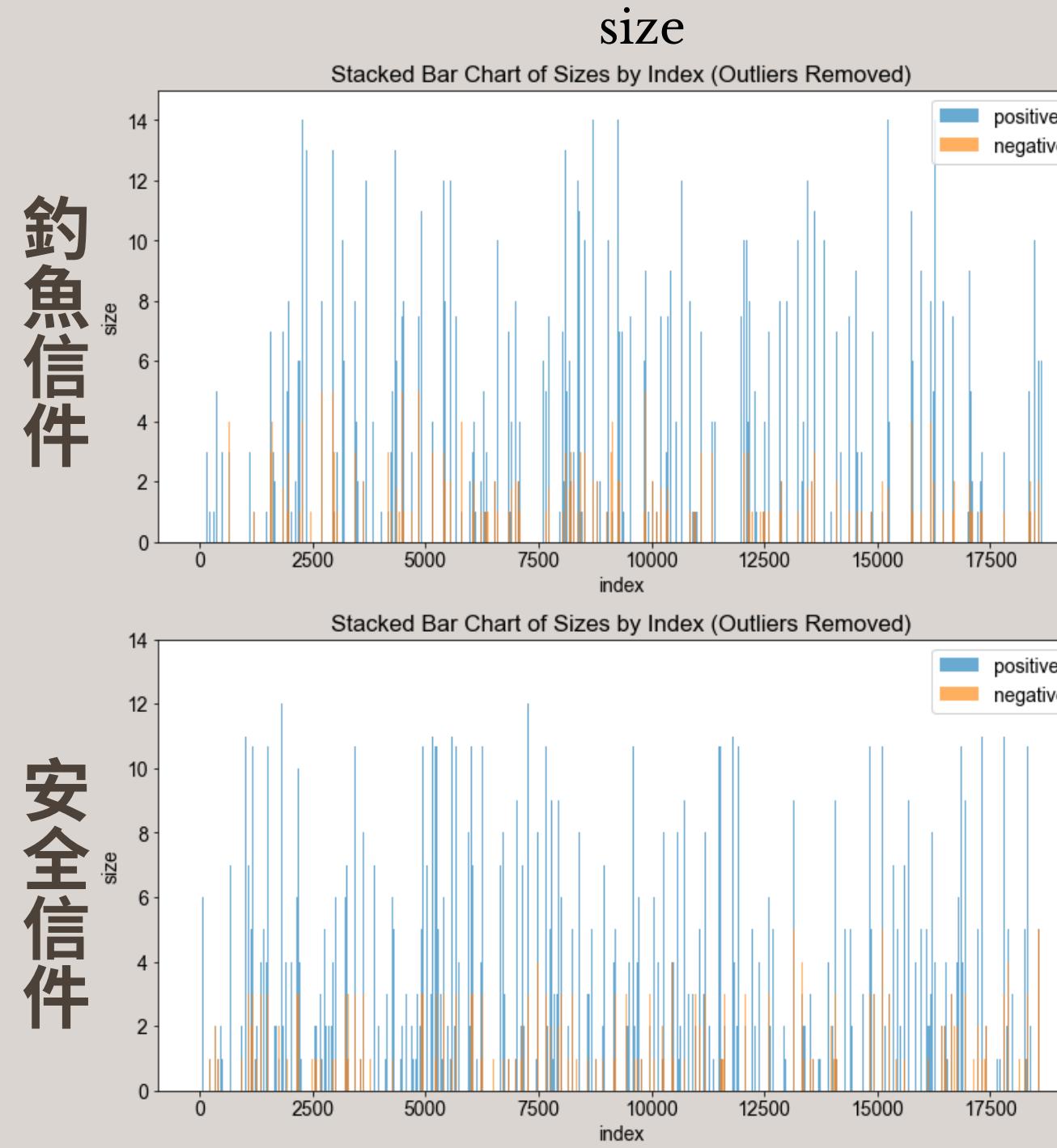


結果顯示情緒偏正向，釣魚信件的正向情緒中位數比安全信件來得高。

Lexicon

情緒直方圖（去除離群值）

因這兩個dataset中有一些離群值致使圖表呈現效果不佳

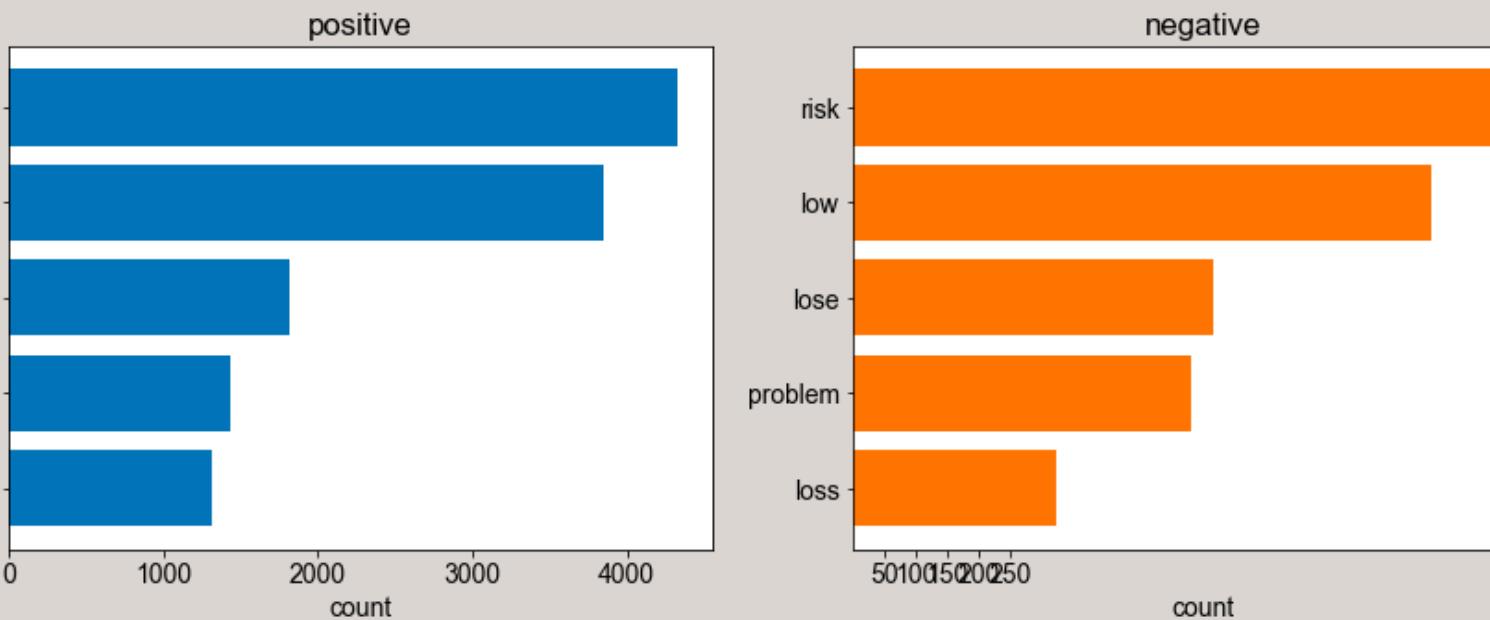


由圖表結果可看出

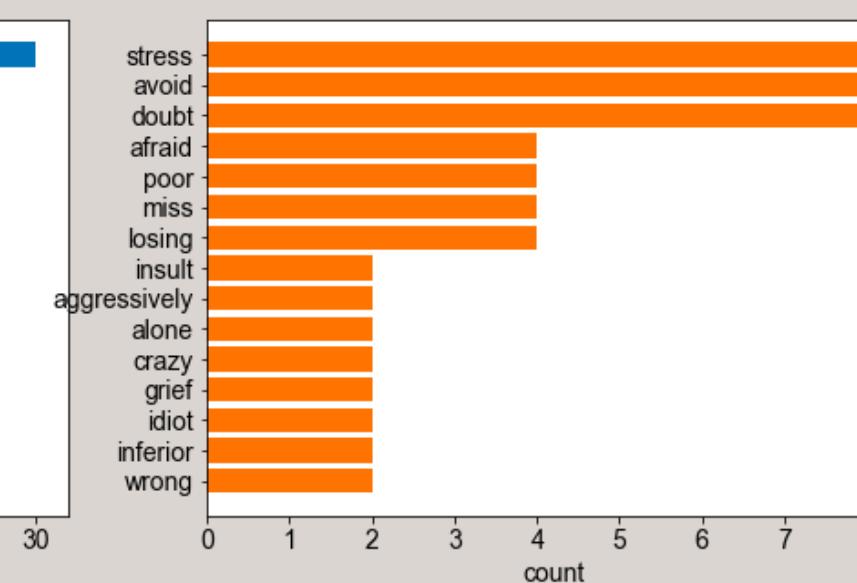
- 兩者分佈之y軸範圍差不多
 - 兩者都顯示幾乎每封信正面詞彙多於負面。

情緒代表字

釣魚信件



釣魚信件(no.14275)



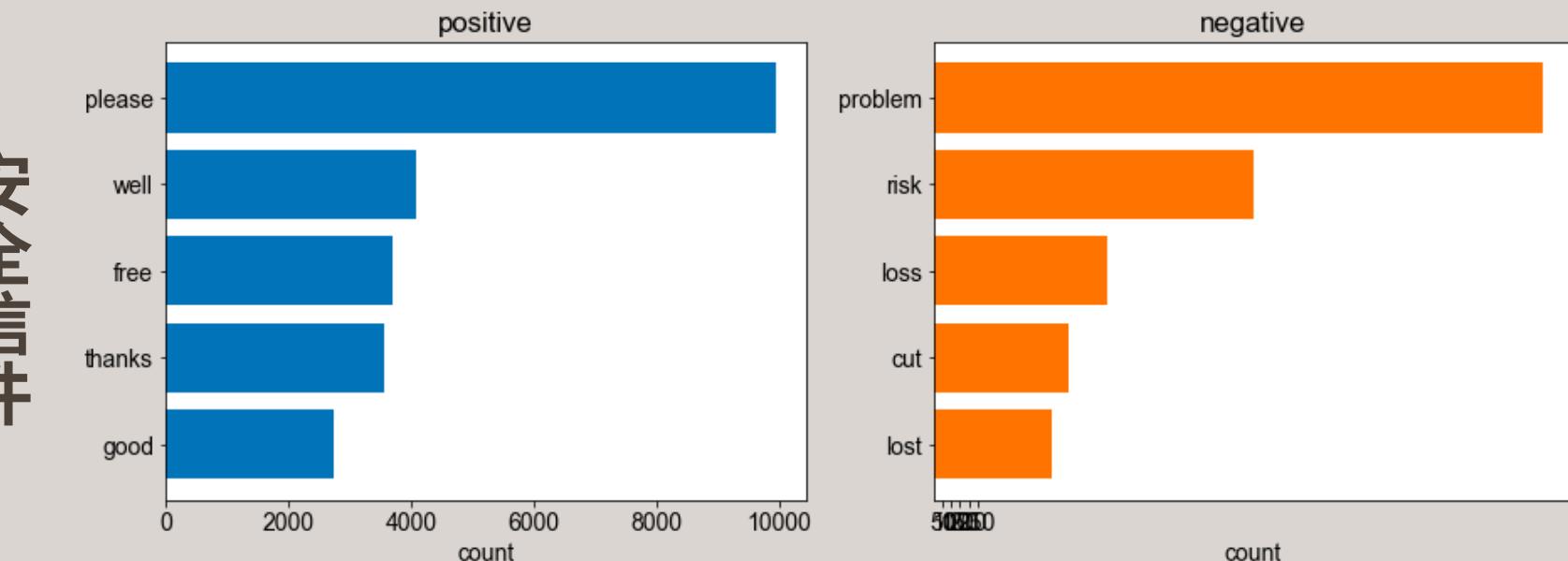
由圖表結果可看出：

釣魚信件的正面詞彙**free, benefit, bonus**比安全信件的**please, well**更加吸引人，釣魚信件可能是利用人類的貪念誘騙上當。

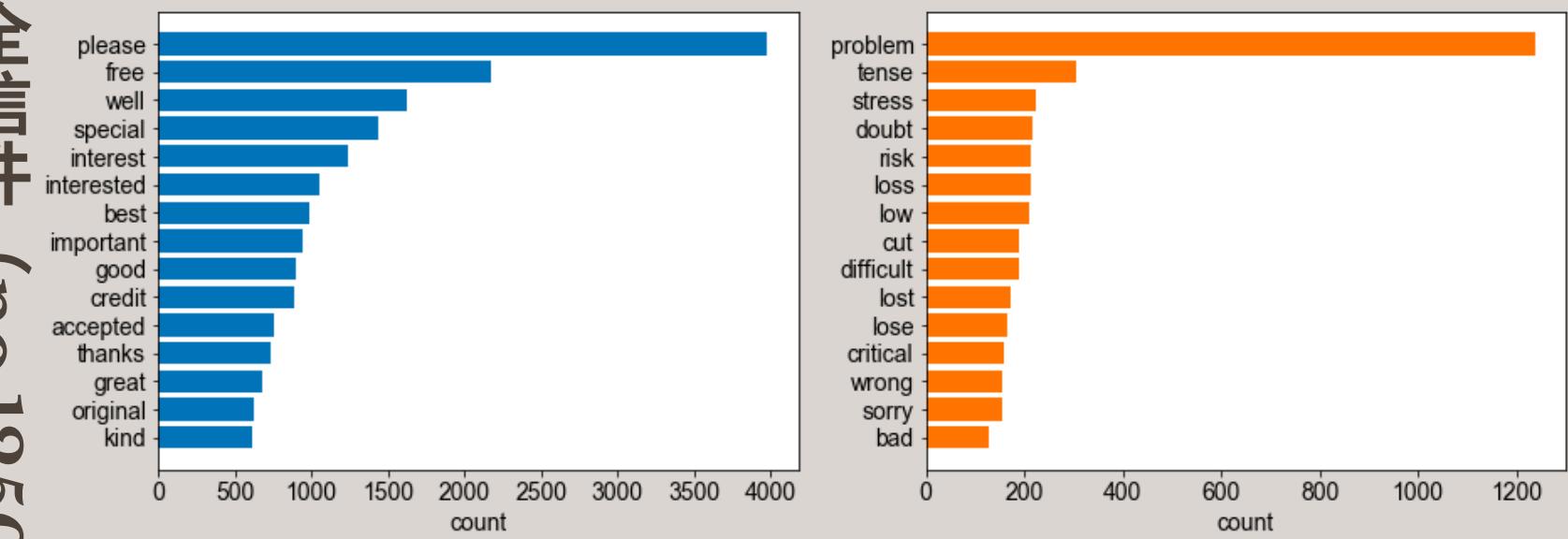
而在負面詞彙中，釣魚信件用了相較於安全信件負面詞彙第一名的**problem**情緒更強烈的**risk, stress**，讓人更擔心受怕。

Lexicon

安全信件



安全信件 (no.12501)



情緒分數

使用 Huggingface 上面已經針對 Sentiment classification 任務 finetune 的 BERT 模型來實作

釣魚信件

Star

- 1: Semi-negation
- 2: Negation
- 3: Neutral
- 4: Semi-positive
- 5: Positive

		sentence	label	score
6974		Now you can have HUNDREDS of lenders compete f...	star 5	0.472031
15453		PUBLIC ANNOUNCEMENTThe new NAME domain extensi...	star 4	0.534208
3275		get the best price on your next car exclusive...	star 5	0.446151
6008		santa barbara nexus ezine great article perta...	star 4	0.533679
13009		Big and bigMAIN PAGEHuge big titties bigbigsc...	star 4	0.565446
14996		fwd finally a smart sp m control solution y...	star 5	0.395988
16620		buy office xp for fifty bucks percentage htmlh...	star 5	0.301693
14334		Â Â Hi Jm Thanks ...	star 1	0.378892
7684		you don t know how to attract customers to y...	star 5	0.424770
13502		work at home a month earn extra income from...	star 5	0.344362
15832		gb q want to establish the office...	star 5	0.297903

- 大多數是含有正面情緒的。

Bert

Langchain Tagging

名稱: google/gemma-1.1-7b-it

總信件取100筆

```
+-----+  
| PromptInput |  
+-----+  
*  
*  
*  
+-----+  
| PromptTemplate |  
+-----+  
*  
*  
*  
+-----+  
| HuggingFaceEndpoint |  
+-----+  
*  
*  
*  
+-----+  
| HuggingFaceEndpointOutput |  
+-----+  
{format_instructions}
```

email_template = """<start_of_turn>user
你是一位網絡安全專家，你將會分析電子郵件的內容，請抓取出郵件中：
1.此郵件的情緒，正向或負向或中性
2.信件的種類，例如釣魚信或安全信
3.信件中提及的關鍵字，並以逗號'，'分隔

以下為一些範例：
```  
範例1  
郵件：我們發現您的賬戶存在可疑活動，請立即登錄確認。  
抓取結果：{{  
"情緒": "負向",  
"信件種類": "釣魚信",  
"關鍵字": "賬戶,可疑活動,登錄"  
}}  
範例2  
郵件：本週五將舉行部門聚餐，請大家提前報名。  
抓取結果：{{  
"情緒": "正向",  
"信件種類": "安全信",  
"關鍵字": "聚餐,報名"  
}}  
範例3  
郵件：您的包裹即將送達，請確認收貨地址。  
抓取結果：{{  
"情緒": "中性",  
"信件種類": "安全信",  
"關鍵字": "包裹,收貨地址"  
}}

## output

|       | index | Email Text                                        | Email Type     | sentence                                          | LLM sentiment                                     |
|-------|-------|---------------------------------------------------|----------------|---------------------------------------------------|---------------------------------------------------|
| 6974  | 7192  | \nNow you can have HUNDREDS of lenders compete... | Phishing Email | Now you can have HUNDREDS of lenders compete f... | {'情緒': '正向', '信件種類': '釣魚信', '關鍵字': '貸款,再入款,房屋貸... |
| 15453 | 15914 | PUBLIC ANNOUNCEMENT:The new .NAME domain exten... | Phishing Email | PUBLIC ANNOUNCEMENTThe new NAME domain extensi... | {'情緒': '正向', '信件種類': '新聞稿', '關鍵字': 'NAME域,個人域名... |
| 3275  | 3386  | get the best price on your next car ! exclusiv... | Phishing Email | get the best price on your next car exclusive...  | {'情緒': '正向', '信件種類': '推廣信', '關鍵字': '車,優惠,報價,買車... |
| 4186  | 4333  | On Thu, 1 Aug 2002 17:10:48 +0100, John Hinsle... | Safe Email     | On Thu Aug John Hinsley wrote No the prob...      | {'情緒': '正向', '信件種類': '安全信', '關鍵字': 'drivers,Me... |
| 6008  | 6198  | santa barbara nexus ezine – great article pert... | Phishing Email | santa barbara nexus ezine great article perta...  | {'情緒': '正向', '信件種類': '安全信', '關鍵字': 'santa bar...  |

```
sentiment
正向 57
中性 22
負向 20
Name: count, dtype: int64
```

graph

template

- 有些抓出的信件種類與原始資料不太相同。
- 情緒的部分大多有抓正確，而LLM是利用關鍵字判別信件種類。也有可能是模型大小的問題致使。

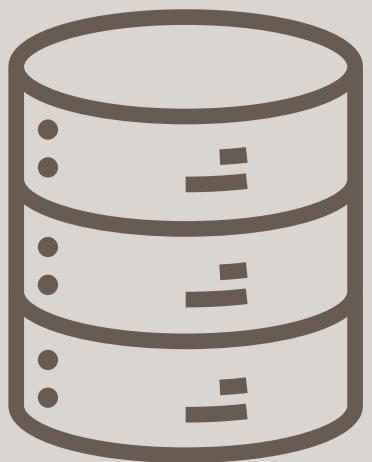
LLM

# Sentiment analysis conclusion

比起安全信件，釣魚信件....

- 整體的情緒較為正面。
- 在用字遣詞上，無論是正面或是負面詞彙都是情緒更強烈的字眼。

# Splitting Dataset



**training set**

13416筆



**validation set**

1491筆

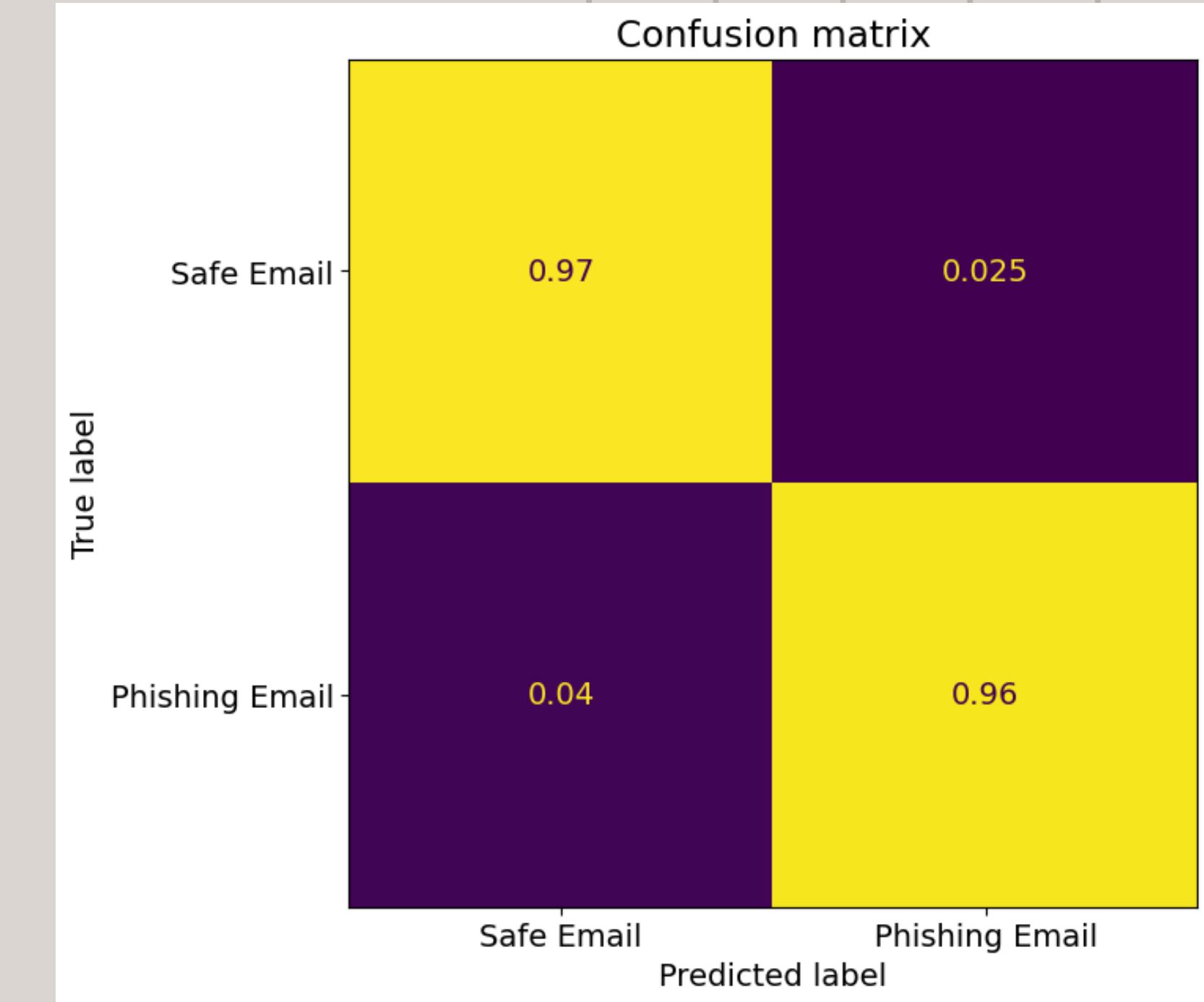


**testing set**

3727筆

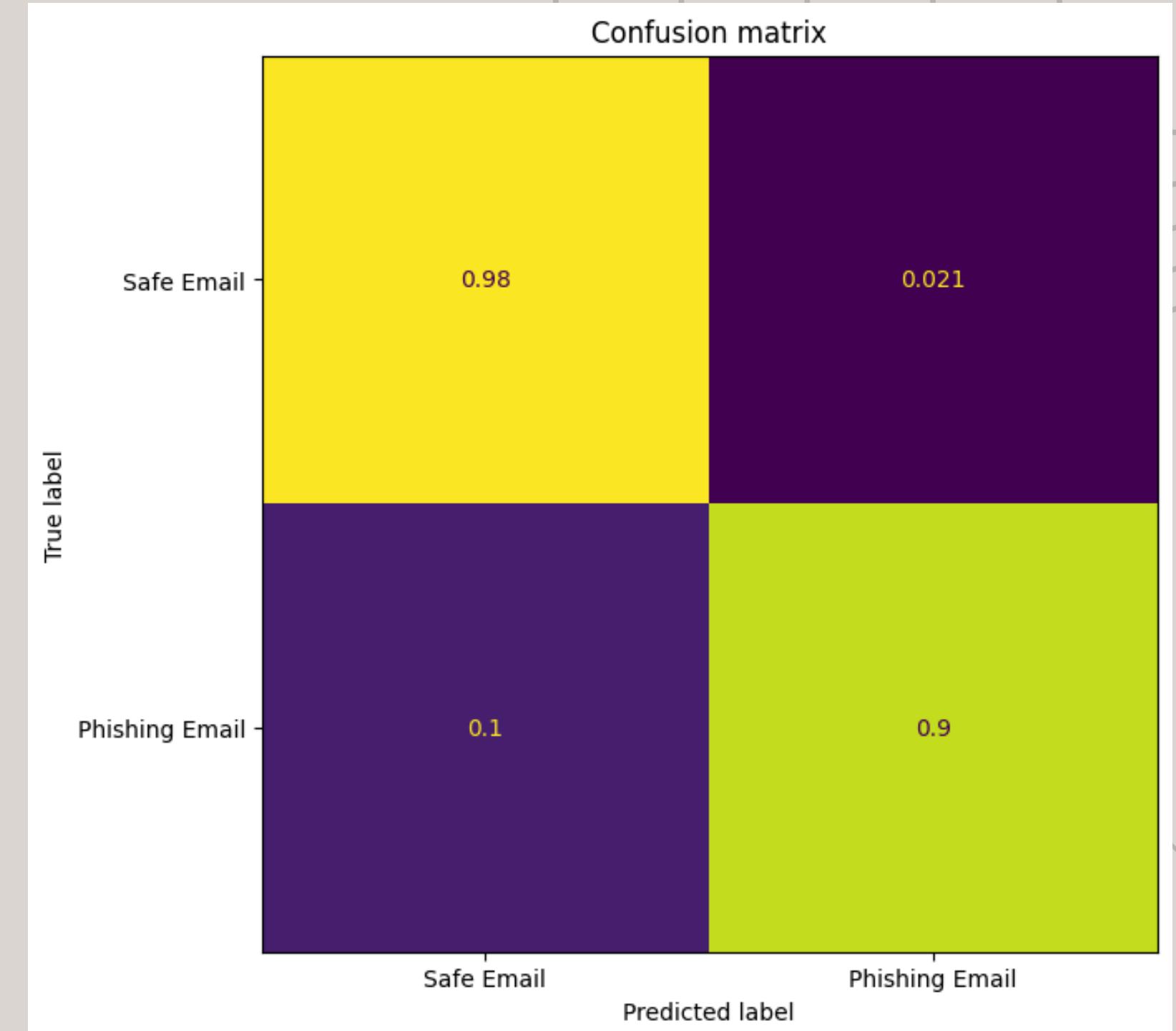
# TF-IDF + Random Forest

|           |      |
|-----------|------|
| accuracy  | 0.97 |
| precision | 0.97 |
| recall    | 0.97 |
| F1-score  | 0.97 |



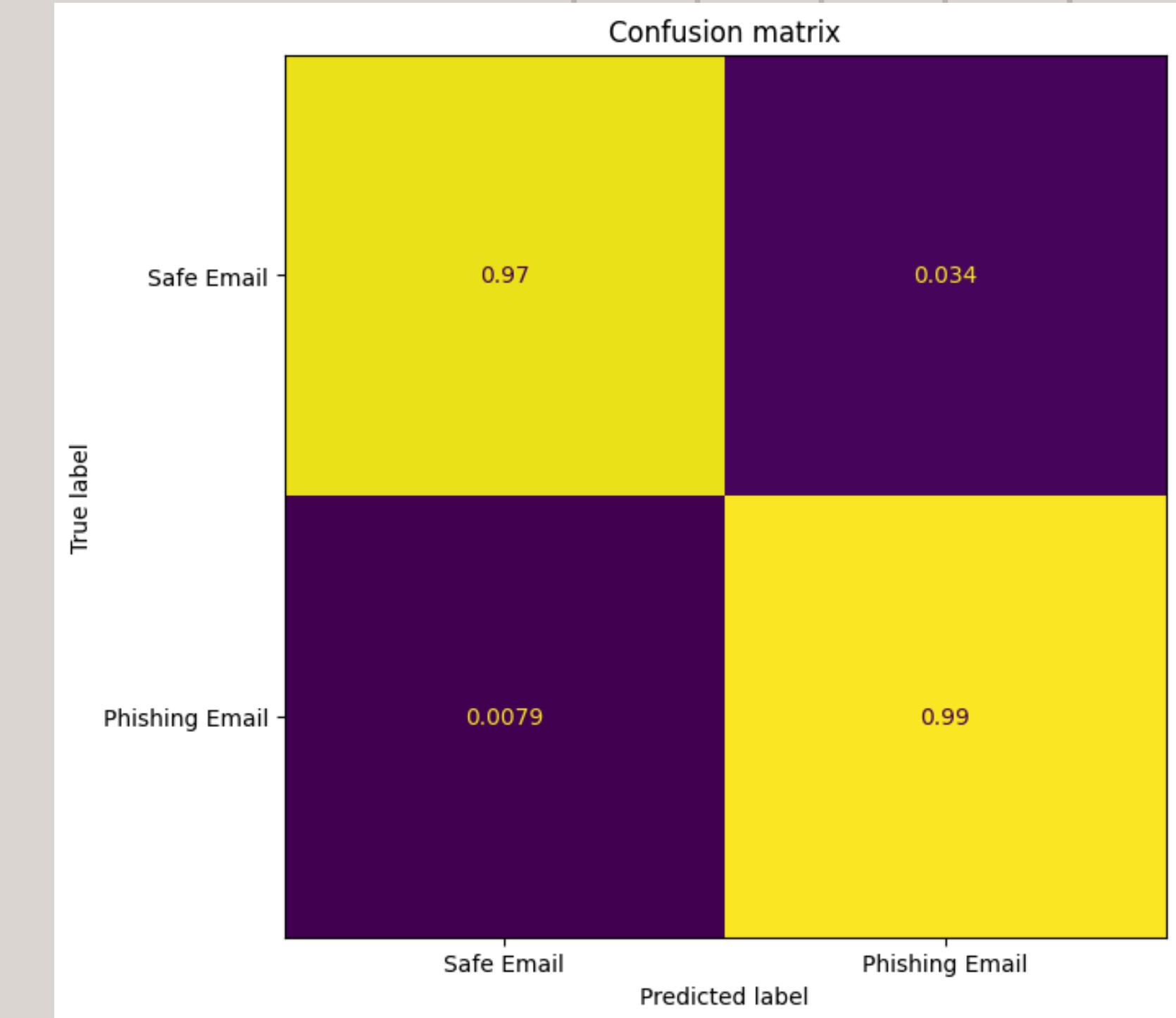
# DistilRoBERTa Embedding + Random Forest

|           |      |
|-----------|------|
| accuracy  | 0.95 |
| precision | 0.95 |
| recall    | 0.94 |
| F1-score  | 0.94 |



# Fine-tune RoBERTa

|           |      |
|-----------|------|
| accuracy  | 0.98 |
| precision | 0.97 |
| recall    | 0.98 |
| F1-score  | 0.97 |



# 解釋性圖表

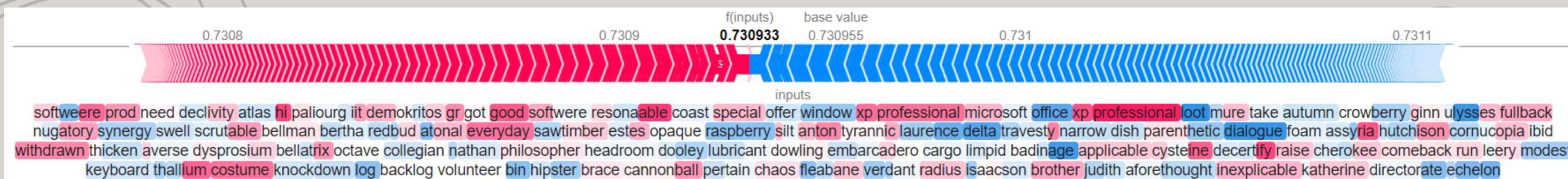
解釋分類器的預測如何形成

## Multiple instance text plot

- 根據 NER 結果：釣魚信件特徵—軟體公司（如：microsoft）
- 取 Email Text 中含有 microsoft 的信件預測
- 預測結果皆為 Phishing Email
- 解釋圖

原文

softweere prod need declivity atlas hi paliourg iit demokritos gr got good softwere resonable coast special offer window xp professional microsoft office xp professional  
loot mure take autumn crowberry ginn ulysses fullback nugatory synergy swell scrutable bellman bertha redbud atonal everyday sawtimber estes opaque raspberry silt  
anton tyrannic laurence delta travesty narrow dish parenthctic dialogue foam assyria hutchison cornucopia ibid withdrawn thicken averse dysprosium bellatrix octave  
collegian nathan philosopher headroom dooley lubricant dowling embarcadero cargo limpid badinage applicable cysteine decertify raise cherokee comeback run leery  
modest keyboard thallium costume knockdown log backlog volunteer bin hipster brace cannonball pertain chaos fleabane verdant radius isaacson brother judith  
aforethought inexplicable katherine directorate echelon



# 總結



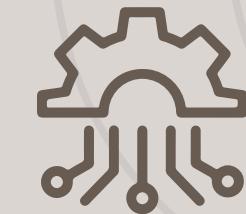
## 關於資料集

- 釣魚信件的情緒：中性偏正向
- 釣魚信件的主題
  - a. 免費軟體
  - b. 金融投資
  - c. 醫療藥品
  - d. 紿予金錢



## 釣魚信特徵

- 釣魚信具明顯特徵
- 軟體公司、金錢優惠
- 正面形容：free, bonus...
- 負面形容：risk, stress...
- 以最高級形容詞激勵使用者  
點擊連結：best, largest...



## 分類器成效

- 效果最佳：  
Fine-tune RoBERTa
- 未來可將生活中真實的 Email  
加入 Train 中，提升泛化能力

悲傷 Peter 與他的快樂小伙伴

Thank You  
For Your Listening