富邦金控 Fubon Financial

Natural Language Processing with Disaster Tweets

Interviewee : 邱亮茗

Date : 2025/04/08

大綱

- 命題挑選之緣由
- Abstract
- 資料集介紹
- 建模內容與成效說明
  - TF-IDF + Logistic Regression
  - TF-IDF + Decision Tree
  - Llama embedding + Fully Connected
  - BERT + Fine Tune
  - Performance
  - Next Step （Data Augmentation）
- 分析過程或使用不同方法所得到之 insight 發現
- 本次數據分析於金融業可能的應用場景
  - 風險控制
  - 促進消費

命題挑選之緣由

本次選擇題目為『Natural Language Processing with Disaster Tweets』，主要基於以下考量：

1. 該題目有助於本公司在各地災害初期做出初略判斷：
   - 調整資產配置
   - 進行地區性業務風險評估
2. 此分類模型亦可延伸應用至其他金融場景
   - 金融詐騙偵測
   - 信用風險評估

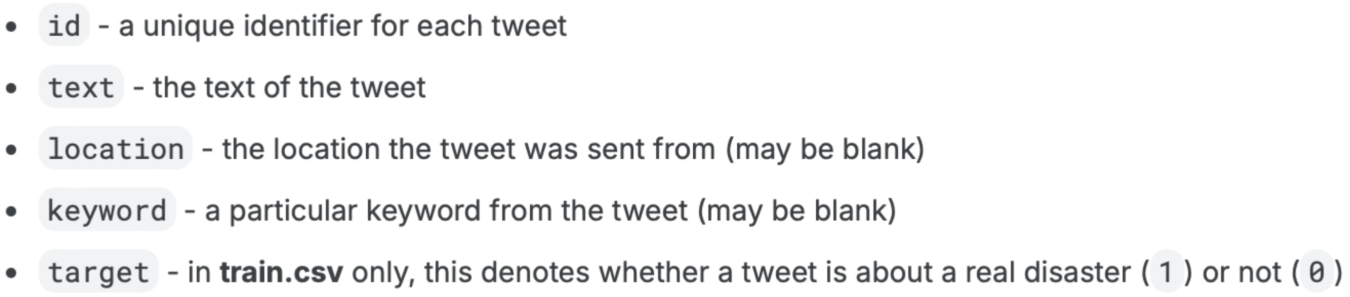
Abstract

This study looks at how Natural Language Processing (NLP) can be used to classify tweets related to disasters, aiming to identify which messages are truly linked to disaster events. The study uses different machine learning models, including TF-IDF with logistic regression and decision trees, Llama embeddings with fully connected layers, and fine-tuned BERT models, to test which methods are best at accurately detecting disaster-related tweets. The results show that deep learning models like fine-tuned BERT and Llama embeddings perform better than traditional machine learning methods in terms of accuracy on both test and private datasets.

This ability to detect disaster-related content from social media allows for timely risk assessments in sectors like finance, where quick responses to emerging threats are important for managing assets and reducing risks. The research also shows the potential of using NLP tools in real-time decision-making to improve response strategies in both finance and disaster management.

資料集介紹

1. Columns



- `id` - a unique identifier for each tweet
- `text` - the text of the tweet
- `location` - the location the tweet was sent from (may be blank)
- `keyword` - a particular keyword from the tweet (may be blank)
- `target` - in **train.csv** only, this denotes whether a tweet is about a real disaster (`1`) or not (`0`)

2. train.csv (7,613)

id	keyword	location	text	target
1	earthquake	California	"Forest fire near La Ronge..."	1
2	NaN	NaN	"Me after watching The Walking..."	0

3. test.csv (3,263)

id	keyword	location	text
0	NaN	NaN	"Just happened a terrible car crash"
2	explosion	NaN	"Heard about #earthquake is different cities, stay safe everyone."

Term Frequency - Inverse Document Frequency

1. Tokenization: 例如 “flooded with people” 轉為 ["flooded", "with", "people"]

2. 刪除常見字: like in, the, with等。例如 ["flooded", "people"]

3. 建詞彙表: 所有出現在 tweets 的單字給一個欄位

4. 計算 TF-IDF 分數：
   - Tweets: "flood flood water everywhere”
   - TF("flood") = 2 / 4 = 0.5
   - IDF(word) = log(總文件數 / (包含該字的文件數 + 1))
   - TF-IDF = TF(word) × IDF(word)

5. 回傳向量: 每個 tweets 轉為 [0, 0.35, 0.71, 0, ..., 0.12] 的向量

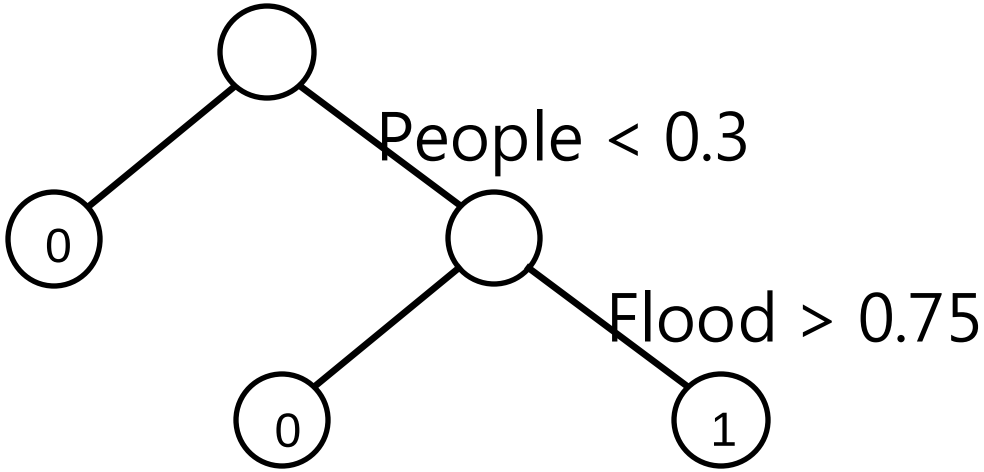
Logistic Regression & Decision Tree

Logistic Regression:

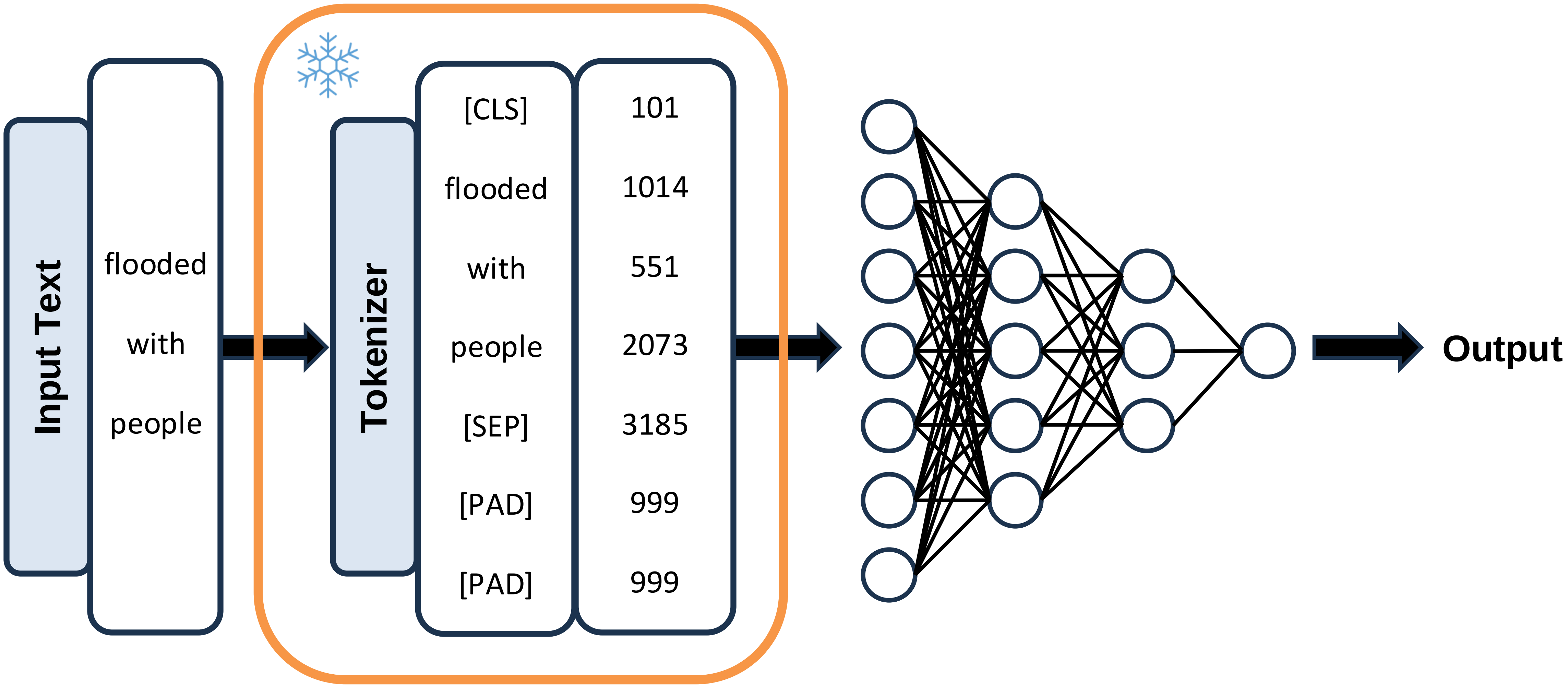
$$z = w_1 \cdot x_1 + w_2 \cdot x_2 + \dots + w_n \cdot x_n + b$$

$$p = \mathrm{sigmoid}(z) = \frac{1}{1+e^{-z}}$$

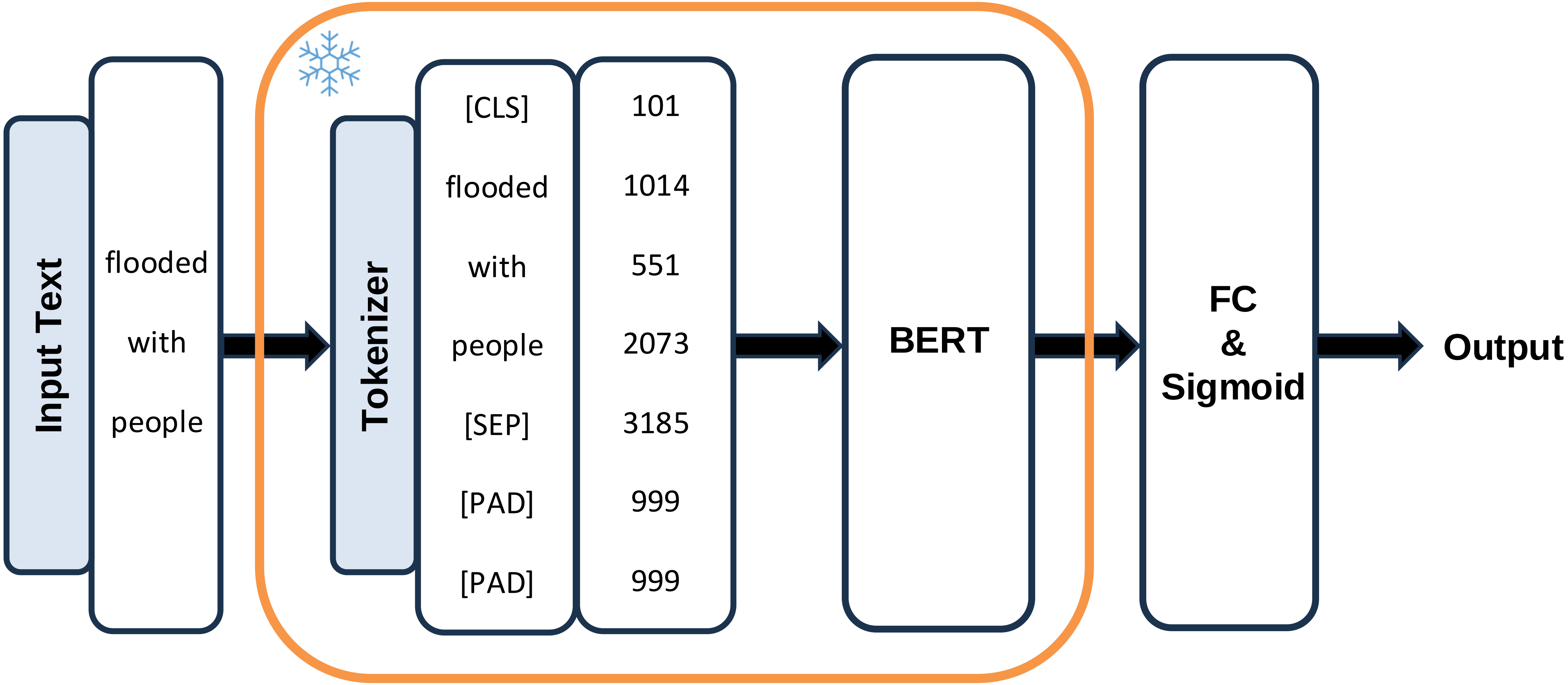
Decision Tree:



Llama Embedding + Fully Connected Layer



BERT + Fine Tune

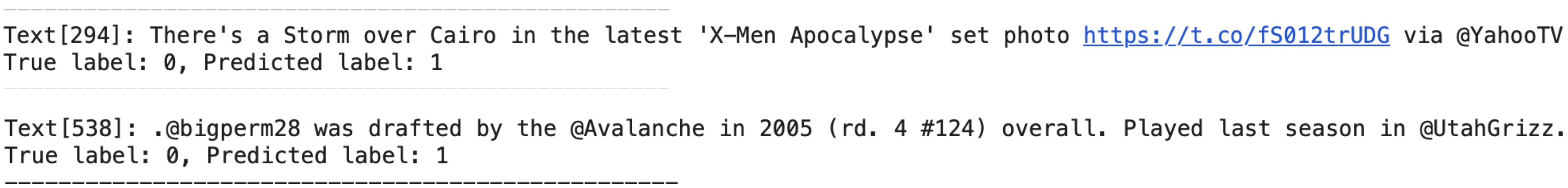


Performance

Model	Accuracy (test)	Accuracy (private)
TF-IDF + Logistic Regression	0.7498	0.7888
TF-IDF + Decision Tree	0.7846	0.7450
Llama embedding + Fully Connected layer	0.8098	0.8194
BERT + Fine Tune	0.8126	0.8207

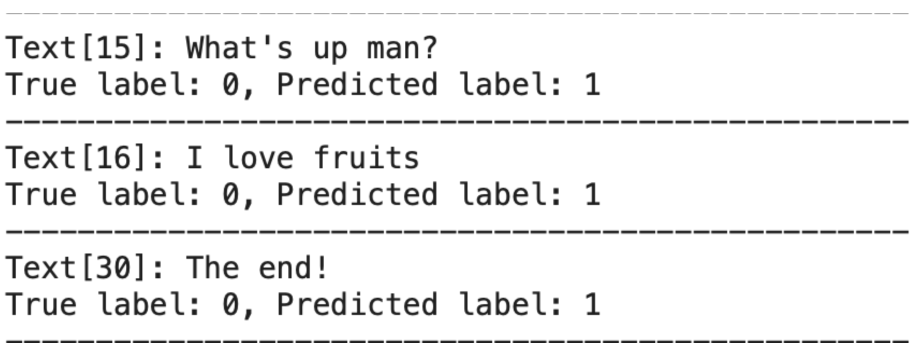
Next Step （Data Augmentation）

- 增強有關專有名詞 (如電影、球隊、歌曲等) 資料訓練。



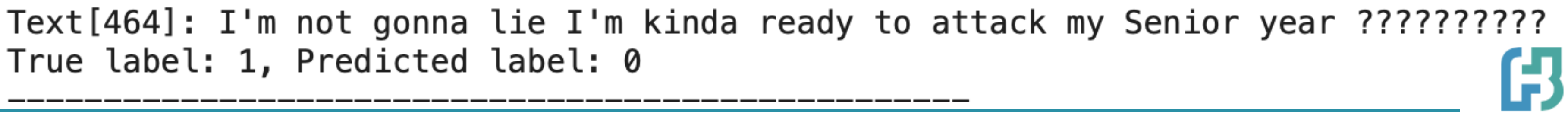
```
--------------------------------------------------
Text[294]: There's a Storm over Cairo in the latest 'X-Men Apocalypse' set photo https://t.co/fS012trUDG via @YahooTV
True label: 0, Predicted label: 1
--------------------------------------------------
Text[538]: .@bigperm28 was drafted by the @Avalanche in 2005 (rd. 4 #124) overall. Played last season in @UtahGrizz.
True label: 0, Predicted label: 1
--------------------------------------------------
```

- 使用Back translation 方式，學習更多語句過短導致誤以為是 disaster 的情境。



```
--------------------------------------------------
Text[15]: What's up man?
True label: 0, Predicted label: 1
--------------------------------------------------
Text[16]: I love fruits
True label: 0, Predicted label: 1
--------------------------------------------------
Text[30]: The end!
True label: 0, Predicted label: 1
--------------------------------------------------
```

- 移除錯誤真實資訊。



```
Text[464]: I'm not gonna lie I'm kinda ready to attack my Senior year ??????????
True label: 1, Predicted label: 0
--------------------------------------------------
```

分析過程或使用不同方法所得到 insight 發現

1. Machine Learning method (TF-IDF + Logistic Regression & TF-IDF + Decision Tree):
   - Logistic Regression 對於資料中的一般語言結構具有穩健泛化能力，但容易受到 TF-IDF 特徵向量稀疏性的限制，無法捕捉語意層次。
   - Decision Tree 容易 overfit 。模型可能記住特定字詞模式，導致在不同分布下表現下滑。
2. Deep Learning method (LLaMA Embedding + Fully Connected & BERT + Fine Tune):
   - 透過 LLaMA embedding 並搭配 Fully Connected layers 的模型在測試集上達到81.94%。
   - 在使用BERT pretrained model 並且經過 fine tuning 後，能夠更好地適應本次的分類任務，進一步提升了模型的準確性。

➢ 傳統 TF-IDF 模型在面對語意模糊、非字面災難用語時準確性受限。

➢ 過往Decision Tree 在部分特徵學習上準確，但容易 overfit，缺乏語境理解。

➢ 大型語言模型（LLaMA、BERT）能有效掌握上下文與語意細節，提升分類準確性。

➢ 使用 pretrained model + Fine-tuning 可以提升在特定任務表現的關鍵策略。

本次數據分析於金融業可能的應用場景 (1)

風險評估

- **預測高風險資產類型:** 根據社群媒體中或新聞報導的訊息，透過分析來識別可能的風險預警信號。例如，若某些行業的消息中頻繁出現負面情緒，這可能預示著該行業的資產將面臨風險，金融機構可以調整貸款政策或資產配置，降低風險。
- **違約風險預測:** 進一步分析"個人"的社群文章內容，結合其過去的行為模式（如消費記錄、社群相片等），可以幫助金融機構提前識別潛在的違約風險。例如，若一個個體在社交媒體上表現出持續的經濟困難或負面情緒，這可能是違約風險的一個警示信號。

本次數據分析於金融業可能的應用場景 (2)

信用卡消費記錄分析

- **消費行為分析：**通過對消費記錄的分析，本公司可以深入了解客戶的消費習慣與偏好。例如，某些客戶可能偏好在線購物、旅遊或餐飲消費，並基於此進行更加精準的產品推薦。
- **個人化廣告推送：**在基於其消費模式推送相關的廣告或促銷活動。舉例來說，若某顧客經常進行戶外用品消費，本公司可推送專屬的產品優惠卡或促銷活動，這不僅能促進顧客消費，也有助於提升顧客忠誠度。
- **客戶細分與定向推廣：**根據消費行為的數據，本公司可以進行客戶細分，從而針對不同客戶群體進行定向推廣。例如，對於年輕客戶，本公司可以推出與生活娛樂相關的優惠活動，而對於高收入群體則可以推送高端商品或服務的廣告。

Thank You!

Q&A

Code availability:
The code repository is stored at https://github.com/R12942159/Fubon_interview.