# Lab2 Kaggle Competition Report

## Summary :

這篇主要講一下我做完 Take Home Exercise 的心得、Kaggle Competition 怎麼做的、以及做這些處理的原因。

## Procedure :

其實一開始在提取資料時有想過要不要利用"hashtags"，後來自己滑了一下後發現 hashtags 比較常標跟推文內容相關的東西，比如地標、或是一些網路詞語之類的，甚至有為了增加流量亂標的，總之跟情感沒甚麼關係、因此捨棄只用"text"。

然後就是 Feature Engineering，我利用 CountVectorizer 將 Training Data 轉換成 BOW、用 One-Hot Encoding 將感情換成向量。之後照抄 Exercise 裡面的 Neuro Network 去訓練。

## Rationale :

主要講兩點，為什麼不用 LLM Embeding 而是 BOW？ 架構為什麼用 Neuro Network 而非其他的？

第一個問題主要是因為在 Exercise 裡面觀察到 LLM Embeding 在 Neuro Network 架構下的表現不太好，因為 LLM 當初就是設計來理解整句話的語意，對於情感預測有太多沒必要的資訊或噪音，所以才選擇 BOW。

第二個問題是為什麼用 Neuro Network，我的想法是因為資料量還蠻多的，用 Neuro Network 應該可以避免 Overfitting，還有現成的 Code 可以用，所以才選 Neuro Network。當然優化的地方也很多，可以試者用其他架構，比如 LSTM / Transformer、也可以試者用 Pre-Trained Word2Vec，雖然可能要研究在 kaggle 如何引用 Pre-Trained Model，可能之後可以多試試看。

## 心得 :

Lab2 Master 我覺得蠻好玩的，實際練習怎麼用 Ollama、RAG 怎麼跑的，對於 LLM 的理解不再只是 Transformer 而已。Homework 的部分有些實作需要研究一下，有遇到 nltk 在 CPU 運行的時候沒事、但用 TPU 加速的時候就會出現問題，或是要研究 Pre-Trained Model 怎麼上傳等。雖然學到就是自己的，但還是覺得有點麻煩就是了。