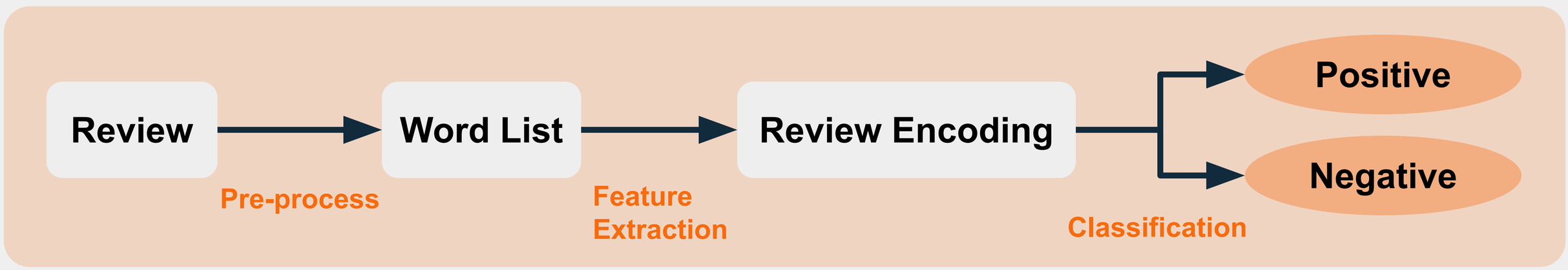
# 電影評論情緒分析

Group 7 : 劉美忻、邱繼賢、張蘞云




## Data pre-processing

Ex : I could n't recommend this film more .

⇒ [‘i’ , ‘could’ , ‘not’ , ‘recommend’ , ‘this’ , ‘film’ , ‘more’]

## Feature extraction

**tf-idf** :

利用詞頻和該詞常用程度將一條評論轉換成向量

**word2vec** :

利用 CBOW 或 skipgram 等方式在考慮前後文語義下，將每個單字轉換成向量

**Bert** :

利用 self-attention 考慮整句評論語義下，將每則 review 轉換成 encoding 向量 C，word tokens 轉換成encoding 向量 T

## Classification model

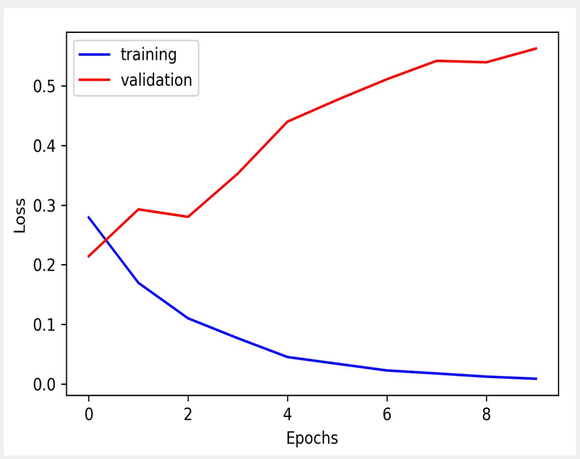
- SVM
- RandomForests (RF)
- Fully conneted neural network (FCN)
- LSTM

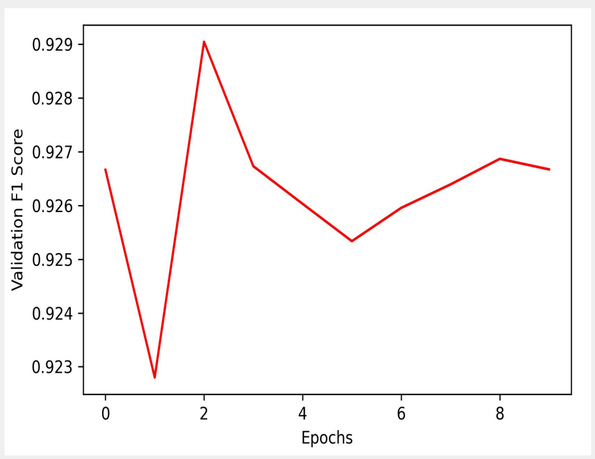
## Our best model : Bert + FCN

- Bert base :

| | |
|---|---|
| Number of layers | L = 12 |
| Size of hidden layers | H = 768 |
| Self-attention heads | A = 12 |
| Total parameters | 110M |

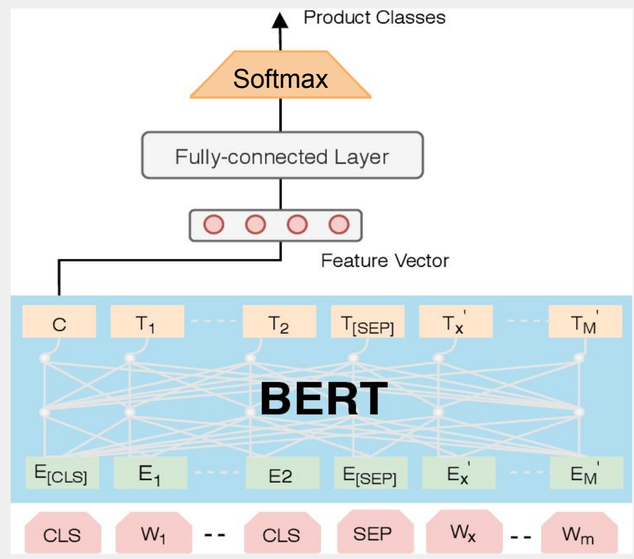
- Loss function : Cross Entopy
- Optimizer : AdamW (learning rate = 1e-5 , eps = 1e-8)
- Batch size = 16 , Epochs = 3
- Training and Validation curve :




## Bert

將每則評論視為 512 個 tokens 輸入進 Bert 模型中，輸出則是 tokens 所各自對應到的 768-dim encodings，處理分類問題時僅需將 [CLS] token 所對應的 encoding C（用以代表整句評論）送進 FCN 進行分類預測。(Pre-trained BERT allows us to use a powerful deep bidirectional representation of each review, fine-tuned for our task.)


## Performance on test data

| Model | F1 score |
|---|---|
| tf-idf + SVM | 0.8731 |
| tf-idf + FCN | 0.8451 |
| w2v + SVM | 0.8472 |
| w2v + RF | 0.8219 |
| w2v+ LSTM | 0.7935 |
| Bert + FCN | 0.9330 |

## Other idea

因為 Bert 模型通常設定輸入的 token size 上限為 512，低於某些 review 的 token size，故可以將每則 review 切成數個 sentences，分次輸入進 Bert 模型中，再將輸出的 encoding C 取平均後送入 FCN 進行分類。