**機器學習與深度學習期中報告**

**TMU ML Course Team Battle 1**

**組別 : 鎮瀾宮的骰子(Feat. 標太郎水餃)**

**組員 : 黃彥鈞(m947108006)**

**石家宜(m946108003)**

**趙上涵(m946108004)**

1. **資料集**

IMDb電影評論資料

| **Data** | **Label** |
| --- | --- |
| 影評內文 | Pos/ Neg |

1. **競賽結果**

Public : 0.87298 (Rank 3)、Private : 0.86924 (Rank 6)

1. **基礎分類模型**
   * 前處理 : Data Representation : TFI-DF
   * 模型得分 :

| **Models** | **Cross\_validation** | **Private\_scores** | **Public\_scores** |
| --- | --- | --- | --- |
| Multinomial NB | 0.865 | 0.740 | 0.761 |
| kNN | 0.775 | 0.663 | 0.672 |
| Decision Tree | 0.704 | - | - |
| Linear Regression | 0.890 | 0.838 | 0.846 |
| Random Forest | 0.739 | 0.706 | 0.691 |
| SVM (kernel = linear) | 0.896 | 0.846 | 0.859 |
| SVM () |  |  |  |
| XGBoost | 0.808 | 0.874 | 0.864 |

* + 最好的基礎模型 : XGBoost

1. **Dnn & Bert**
   * Dnn
   * Bert

本組採用的為建構於PyTorch框架的Bert模型，並將在原始文本資料經過轉換成Bert接受的tsv檔案，並進行特徵萃取。最後再讀入Pre-trained的Bert模型，進行Tokenization和Fine Tuning後，交付模型訓練。

| **Model** | **Private\_scores** | **Public\_scores** |
| --- | --- | --- |
| bert binary classification (3 epochs) | 0.869 | 0.873 |
| bert binary classification | 0.860 | 0.867 |

1. **Conclusion**