




中文新聞分類與分析系統

吳佳臻



⚙️動機

- 新聞網站每日產出大量內容，類別繁多（政治、生活、國際等）
- 缺乏標準化自動分類，仰賴人工整理，效率低、易出錯

🎯目標

- 建構一套中文新聞自動分類系統
- 提升資訊處理效率，可應用於推薦、輿情監控等場景



資料與前處理

資料來源

- numbda-webnew
- 中文新聞語料

預處理

- 移除 HTML 標籤與特殊符號
- CKIP 斷詞處理
- 對應類別與 label 編碼（共 13 類）

類別分類

政治、財經、國內、國際、品味生活、公民運動等

模型與訓練設定

使用模型

- BERT(bert-base-chinese)
- BERT (WWM) (hfl/chinese-bert-wwm-ext)
- RoBERTa (WWM) (ckiplab/bert-base-chinese-wwm)

訓練設定

- Hugging Face Trainer
- batch size, epochs, learning rate 等參數微調
- 損失函數：
CrossEntropyLoss



分類結果與評估

評估指標

- Accuracy
- Precision / Recall / F1-score
- Confusion Matrix

結果摘要

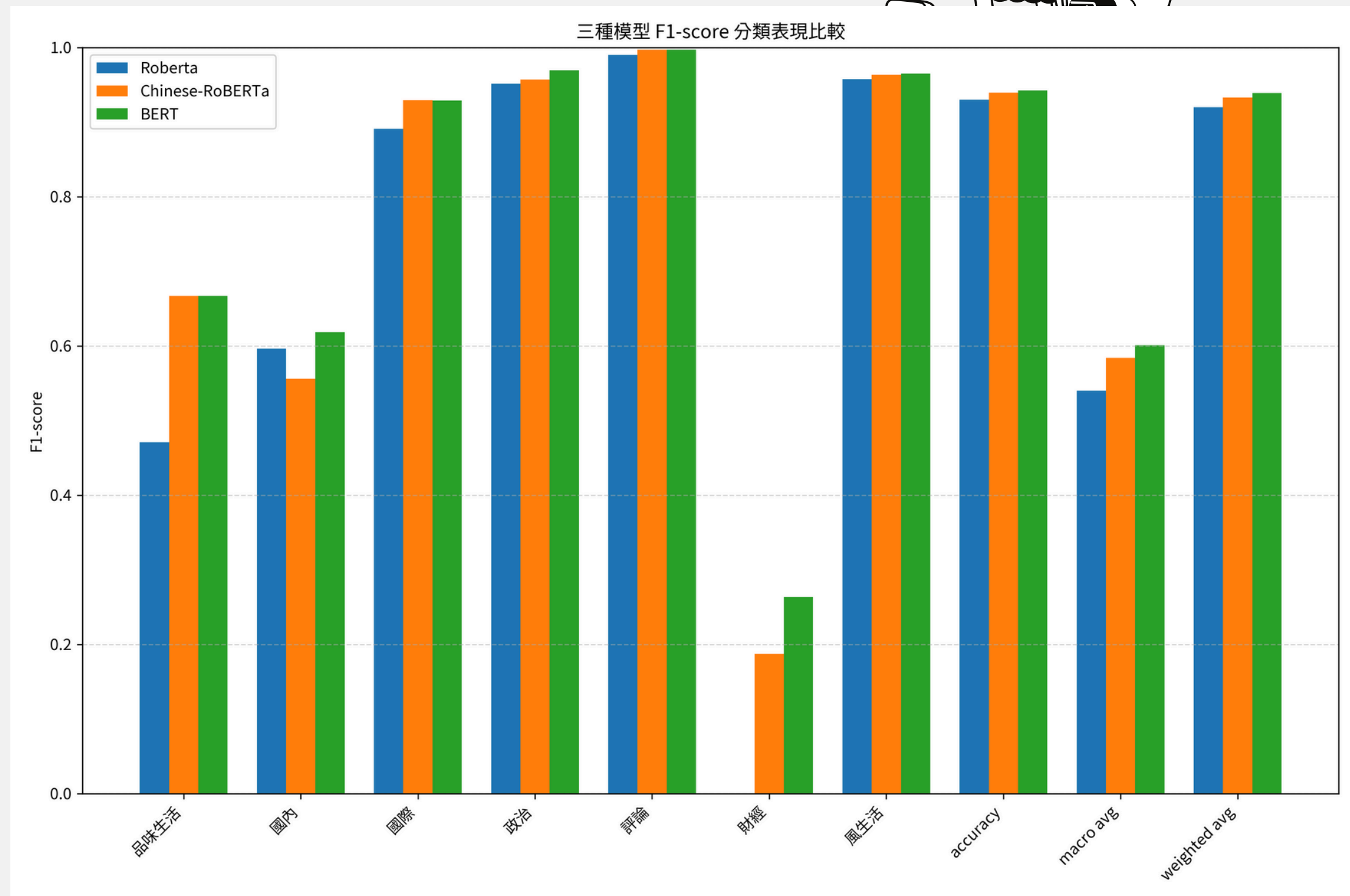
- BERT 在大多數類別 F1-score 表現最佳
- 品味生活類別中，BERT 與 Chinese-RoBERTa 表現相當
- 中港澳、公民運動等少數樣本類別無法有效分類 ($F1 = 0$)



分類結果與評估

評估指標

- Accuracy
- Precision / Recall / F1-score
- Confusion Matrix



應用場景與展望

潛在應用

- 新聞平台自動分類
- 使用者個人化推薦
- 公部門輿情分析工具

未來展望

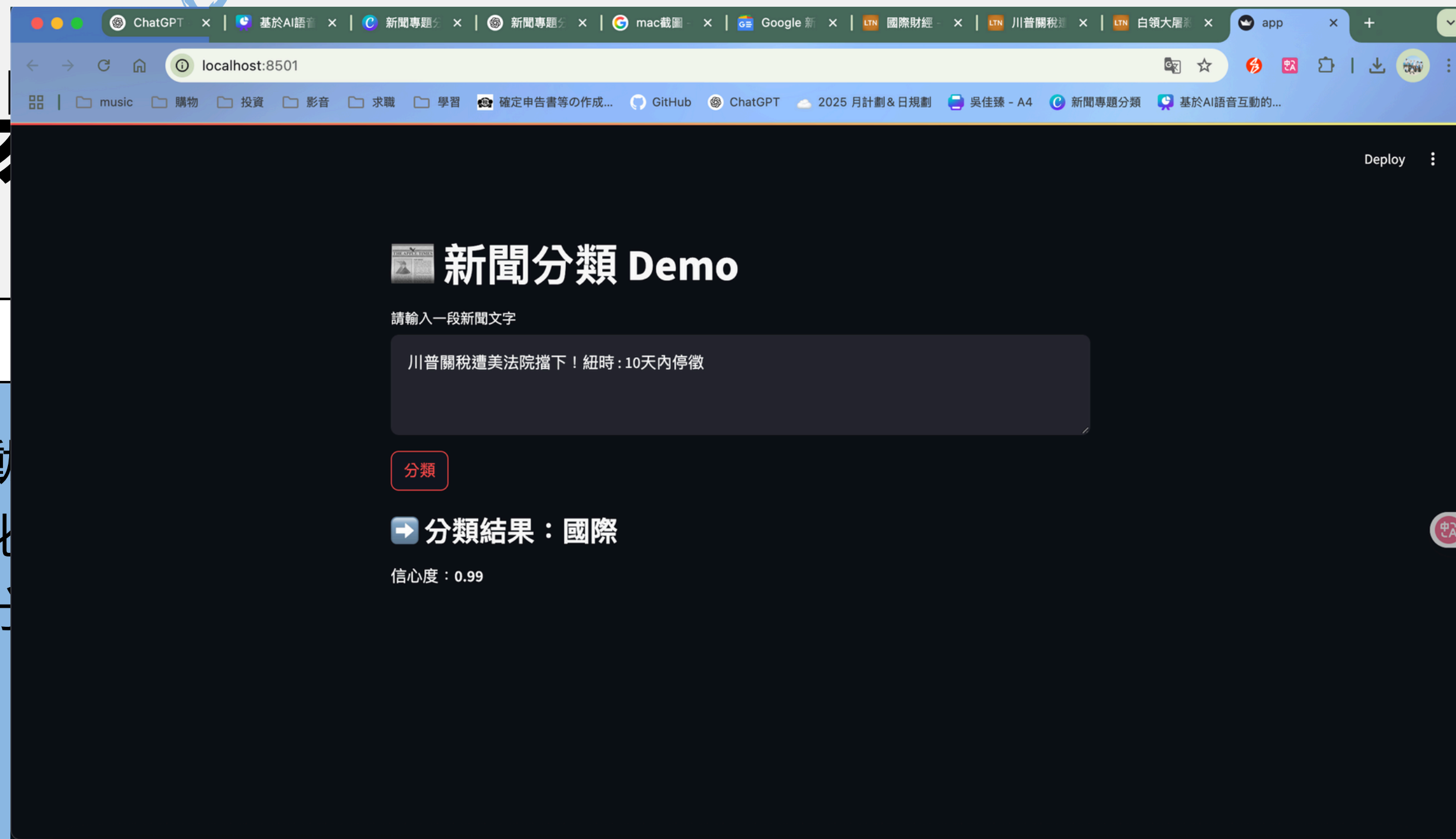
- 加入多模型融合投票機制
- 支援摘要與情感分析



應用場景

潛在應用

- 新聞平台自動分類
- 使用者個人化推薦
- 公部門輿情分析





結語

- 成功建構中文新聞分類模型，完成分類、評估與錯誤分析
- 展望未來應用於新聞推薦與輿情監控等場景

 附件

<https://github.com/jenny870604/news>

The background of the slide is decorated with various hand-drawn blue scribbles and shapes. These include loops, swirls, zig-zags, and abstract patterns that frame the central text. The style is casual and artistic, resembling a child's drawing or a hand-painted background.

**Thank you
very much!**