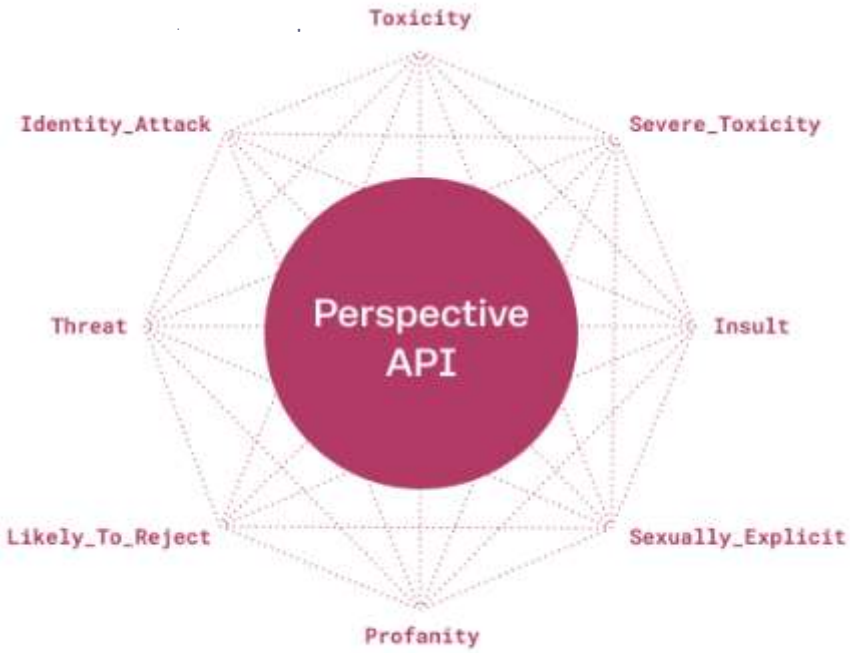
自然語言處理毒文分析

指導教授：余瑞琳老師
專題學生：吳晉慧、陳宜妤、游惠晴、葉素鳳、謝宜娠、段雁庭、林庭均

動機和目的

- 自從社群媒體時代興起以來，網路上有許多人會發表自己的意見，其中大部分都有大量的消極性和毒性。而有毒內容可能會對心理健康產生不利影響。因此我們透過網路霸凌和毒性的資料集分析社群媒體上的相關評論進行實驗。

動機和目的



我們的研究會透過機器學習、神經網路等方法，來進行有毒評論的分析，並擴展分析評論是否涉及到網路霸凌。

輸入評論 → 分析是否為網路霸凌 → 分析包含之毒性 → 輸出標籤

資料介紹

- 有毒評論資料集：分析Twitter上的毒性評論
- Google的內部部門Jigsaw開放的有毒評論挑戰(2017開放)

特徵
有毒的
嚴重有毒的
猥褻毒性
威脅毒性
侮辱毒性
身分仇恨毒性

- 網路霸凌資料集：分析Twitter上的網路霸凌評論
- 年IEEE 國際大數據會議(IEEE BigData 2020) 會議記錄。

特徵	
Age	年齡霸凌
Ethnicity	族群霸凌
Gender	性別霸凌
Religion	宗教霸凌
Other type of cyberbullying	其他類型的網路霸凌
Not cyberbullying	不是網路霸凌

資料介紹

- **數據清洗**

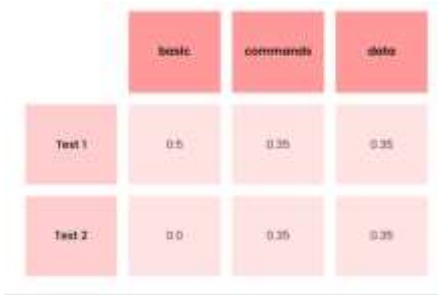
- Twitter上會有很多表情符號
- 英文縮寫對於模型理解會造成偏差
- 因為目前用的語言模型是預訓練在一般維基百科的正式文章上，無法理解網路用語，所以我們需要把俚語展開成完整敘述

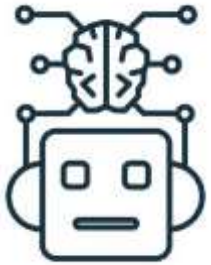
切分 Training, Validation, Testing Dataset，比例 7 : 2 : 1

方法

- 流程

輸入有毒評論 → 特徵提取 → RandomForest / SVM / KNN / LSTM → 輸出





努力才能成功 ⟶ 努力 才 能 成功

他的領導才能很突出 ⟶ 他 的 領導 才能 很 突出

方法

- **網路霸凌**
  - 1. 使用 TFIDF + 隨機森林
  - 2. 使用Word2Vec + LSTM
  - 3.神經網路分類器使用自然語言模型 BERT，加上向量的特徵進行效能分析

- **毒性**
  - 1. 使用 TFIDF + 隨機森林
  - 2. 使用 TFIDF + 支援向量機
  - 3. 使用TFIDF + KNN
  - 4.神經網路分類器使用自然語言模型 BERT，加上向量的特徵進行效能分析

接下來有不雅字眼，請各位做好心理準備

危險

(警 50)

- 實驗結果(網路霸凌+毒性)

#性別霸凌
#有毒的、#猥褻毒性、#侮辱毒性

#其他類型的網路霸凌
#有毒的

#宗教霸凌
#有毒的、#侮辱毒性、#身分仇恨毒性

05 結論

結論

- 基於Google使用機器學習來識別有毒語言，我們還透過機器學習和神經網路的方法進行網路霸凌加毒性檢測的評論研究，而在兩種研究中的方法準確率都高達85%。
- 在本研究中，我們採用了不同的實驗方法，分別是TFIDF搭配隨機森林、TFIDF搭配支持向量機（SVM）、TFIDF搭配K最近鄰（KNN）、Word2Vec搭配LSTM以及BERT模型，其中使用BERT模型效能最好。
- 未來的研究方向可以進一步優化BERT模型，使其在識別有毒評論方面更加準確和高效。此外，擴展研究對於不同語言和文化背景下的網路霸凌檢測，以及應用於不同的社交媒體平台上。這些努力將有助於建立更安全、更友善的線上交流環境，保障用戶的心理健康和個人尊嚴。

THANK YOU