全華
教育新知月刊
2021.02 第017號

想知道更多資訊
請前往全華
大專資源網

資料出處：
iThome https://www.ithome.com.tw/news/142740

臉書強化AI技術，大幅提升騷擾與仇恨言論偵測效果

臉書的人工智慧工具現在能綜合分析文字、圖片和其他細節，考量整體貼文背景資訊後辨識出惡意評論

臉書更新社群標準執行報告，在2020年第4季，騷擾與仇恨言論有明顯下降趨勢：

1. 仇恨言論的普及率下降到0.07-0.08%
2. 暴力內容下降到0.05%
3. 成人裸露內容下降到0.03-0.04%

強化AI技術的人工智慧工具：

1. 對貼文的**圖像**、**文字**和**其他細節**進行組合分析。
2. 建構出**擅長分析評論**，並且能不斷**從新資料學習的系統**。
3. 大幅增進臉書**偵測霸凌和騷擾文字**的能力，並主動刪除這些內容。
4. 支援多種語言，例如可偵測**西班牙語**和**阿拉伯語**等違反政策的內容。
5. **仇恨語音**內容偵測能力也大幅提升。