借助 AI 能更迅速識別社群媒體中的仇恨與恐嚇

駐瑞典代表處教育組

如今，社群媒體中仇恨與恐嚇已成為我們的日常，然而，大量的發表頻道輕易讓人發言，使頻道管理者和警方更難監管這些言論的內容。因此，烏普薩拉大學心理學研究所所長 Nazar Akrami 及其研究團隊，將研發一種自動偵測工具，以期能快速辨識社群媒體中的仇恨與恐嚇言論。

Nazar Akrami 表示，即將研發的工具能協助警方與媒體公司，在部落格、聊天室等類似的數位媒體中，迅速找出仇恨與恐嚇言論。這項計畫也屬於政府積極查緝極端主義的行動之一。

儘管現今已有自動辨識系統，能找出大約百分之五到十之間的仇恨與恐嚇言論，但效果畢竟不彰。若研發出能自動學習的辨識系統，則可望顯著提高辨識精準度。

即將研發的是一套人工智慧系統，研究團隊正持續為系統建立語言規則，但系統仍需透過研讀大量文章來學習語言，例如，研究團隊已將整套瑞典文的維基百科匯入系統內，好讓系統自行進階學習研究團隊尚未建立的語言規則，系統會自動尋找可能的模組和相似性，或是詞序、或是句子長度，都可能和其他數千種變化一起進行比對。

研究團隊除了教這套系統學習瑞典文之外，也代入帶有仇恨與恐嚇的文章，讓系統找出模式和相似性，好讓系統辨認出這類文章和其他類型文章的不同。其他現存的辨識系統，是靠驗明仇恨與恐嚇字眼的方式來過濾，但用以建立系統的樣本數卻相當有限，因此成效不彰；但透過這套能自行分析文章並找出差異的人工智慧系統，就有希望能明顯提高辨識的精準度。

研究團隊選取的文章來源很多樣，其中一種方式是從不同的部落格提取文章，研究人員親自將大量的恐嚇言論判讀為五個等級；另外，研究團隊也從實際上的「仇恨言論產生器」蒐集了約五萬則仇恨文章，研究人員正親自將這些文章分級。

為了提高這套系統的辨識精準度，研究團隊正積極與媒體公司展開合作，以獲得更多分析的素材，以期提升百分之十至二十的精準度，下一步，則可望由這套人工智慧系統自行搜尋仇恨與恐嚇的文章，再由研究人員手動驗證搜尋結果。

撰稿人/譯稿人：莊郁馨

資料來源：2019 年 6 月 3 日，烏普薩拉大學新聞稿

https://www.uu.se/nyheter-press/pressmeddelanden/pressmeddelande-visning/?id=12708&area=6,10,16,25,46,48&typ=artikel&lng=sv