AI SKUTECZNĄ BRONIĄ W WALCE Z MOWĄ NIENAWIŚCI?

Amerykańscy specjaliści stworzyli model sztucznej inteligencji, który skutecznie wyszukuje w sieci pozytywne komentarze. Innowacyjne rozwiązanie może być skutecznym narzędziem w walce z mową nienawiści.

Badacze z Uniwersytetu Carnegie Mellon w Pitsburgu poinformowali, że opracowali technologię AI, która analizuje komentarze w Internecie i wybiera te, które są pozytywne, np. bronią mniejszości czy sympatyzują z określonymi grupami społecznymi.

Naukowcy przeprowadzili eksperymenty, analizując niemal milion wpisów, umieszczonych w serwisie Youtube i koncentrujących się na dwóch wydarzeniach - kryzysie związanym z uchodźcami Rohingya z Birmy oraz samobójczym ataku terrorystycznym, do którego doszło w miejscowości Pulwama w Kaszmirze w lutym 2019 roku.

Stworzono modele językowe, które wyszukiwały określone wyrażenia stosowane w komentarzach, a także były w stanie pogrupować wyrazy o podobnym znaczeniu i zauważyć, kiedy są one stosowane w różnych komentarzach w podobnym kontekście. AI na podstawie przykładowych treści była też w stanie "nauczyć" się przewidywać, jakie słowa mogą pojawić się w kolejnych wpisach. Ponadto, jak podkreślali twórcy programu, był on w stanie przeanalizować wpisy zamieszczone w językach południowoazjatyckich, co zwykle sprawia trudności, jako że często zawierają one kombinacje różnych dialektów.

Amerykańcy badacze poinformowali, że podczas eksperymentu stworzonemu przez nich modelowi AI udało się znaleźć i wydobyć na powierzchnię 88 proc. pozytywnych komentarzy, w momencie, kiedy zwykłe algorytmy znajdowały jedynie 10 proc. z nich.

Naukowcy mają nadzieję, że dzięki ich technologii uda się stworzyć system, który pozwoli na automatyczne wyszukiwanie i podkreślanie pozytywnych wpisów, zamieszczanych na stronach internetowych czy w mediach społecznościowych. Miałoby to pomóc w walce z mową nienawiści w sieci.

Z raportu przygotowanego przez Ligę Przeciw Zniesławieniom (ADL), amerykańską organizację, której celem jest śledzenie i walka z antysemityzmem, jeden na pięciu internautów spotkał się w sieci z pogróżkami i przemocą. 20 proc. padło ofiarą prześladowania w Internecie ze względu na swoją płeć, rasę, pochodzenie etniczne, orientację seksualną, religię czy niepełnosprawność.