KRZYSZTOF DOMARADZKI

*Algorytmy sztucznej inteligencji mogą pokonać szerzącą się w internecie nienawiść. Ale potrzebują do tego wsparcia polityków i przedsiębiorców*

# TECHNOLOGIA KONTRA HEJT

# *Raszpla! Debil! Pedał!*

FOT.: GETTY IMAGES

Dla Michała Wroczyńskiego, twórcy Samurai Labs, platformy walczącej z przemocą w sieci przy pomocy sztucznej inteligencji, ostatnie tygodnie są wyjątkowo intensywne. Gdynia, Warszawa, San Francisco – ciągle w podróży. Zawieszony jest między budowaniem technologii, szukaniem inwestorów a opowiadaniem w mediach o walce z hejtem. Tylko w ciągu kilku tygodni Wroczyński odwiedził m.in. TVN, Onet, radio RDC i „Rzeczpospolitą".

To znamienne. Po śmierci prezydenta Gdańska Pawła Adamowicza – niezależnie od praprzyczyny jego zabójstwa – w Polsce rozgorzała dyskusja o mowie nienawiści i cybernękaniu. Momentalnie wzrosła świadomość, że przemoc w sieci to realny problem społeczny, który odciska piętno na życiu nastolatków, świecie polityki czy funkcjonowaniu firm. A eksperci – psychologowie, socjologowie, ale też specjaliści od AI – zaczęli w mediach szukać sposobu na rozwiązanie problemu.

**A PROBLEM JEST ZNACZĄCY.** Według fundacji i-SAFE ponad połowa młodzieży była nękana w sieci. Cyberbullying Research Center podaje, że 10 do 20 proc. nastolatków doświadcza tego regularnie. A z analizy Harford County Examiner wynika, że tylko 1 na 10 dzieci przyznaje się rodzicom, iż padło ofiarą cybernękania. Z mową nienawiści jest podobnie. Z przeprowadzonego w 2017 roku sondażu Centrum Badań nad Uprzedzeniami Uniwersytetu Warszawskiego wynika, że tylko połowa Polaków uznaje za zdecydowanie obraźliwe skrajne przykłady mowy nienawiści wobec Żydów, Ukraińców czy uchodźców. Połowa z nas nie potrafi jednoznacznie potępić sformułowania: „A niech uchodźcy sobie przybywają do kraju

*Skoro osłaniamy nasze komputery, dlaczego mielibyśmy nie zapewnić ochrony naszym dzieciom?*

**MICHAŁ WROCZYŃSKI**
twórca Samurai Labs

nad Wisłą. Będzie czym palić w elektrociepłowni". Eksperci nie mają wątpliwości: to epidemia przemocy. Coraz brutalniejsza i złośliwa. Sprytna i potrafiąca się kamuflować. Ale jest grupa ludzi, którzy uważają, że potrafią wynaleźć panaceum. Szukają go w rozwiązaniach bazujących na sztucznej inteligencji – znakomicie analizujących tekst pisany, wychwytujących konteksty wypowiedzi, rozumiejących zależności między rozmówcami, rozszyfrowujących mowę ezopową. Krótko mówiąc: pływających w meandrach językowych równie dobrze jak ludzie.

– Są dziesiątki synonimów oraz innych sposobów maskowania nienawiści skierowanej w stronę różnych grup społecznych. Niektórzy użytkownicy internetu, nie wnikając w ich motywy, wykazują się daleko idącą kreatywnością i wytrwałością w szerzeniu nienawiści – mówi Paweł Godula, dyrektor analityki konsumenckiej w specjalizującej się w sztucznej inteligencji firmie deepsense.ai.

– Napastnik może zaatakować ofiarę, publikując kompromitujący ją post, bez użycia wulgaryzmów czy gróźb. Trzeba rozumieć, jak działa społecznościowy świat, żeby wiedzieć, co tak naprawdę jest obraźliwe – dodaje Gilles Jacobs z Ghent University.

*Ciapate ścierwo!*

To specjalizujący się w Big Data i przetwarzaniu języka naturalnego naukowiec, który stworzył autorski system do wykrywania niewłaściwych treści w sieci. Jacobs i jego wspólnicy przetestowali swoje rozwiązanie na popularnej młodzieżowej platformie społecznościowej AskFM, która na przełomie 2012 i 2013 roku przeżyła poważny kryzys spowodowany serią samobójstw jej nastoletnich użytkowników – po otrzymaniu w serwisie anonimowych wiadomości. System wychwycił ponad dwie trzecie postów zawierających groźby, zniewagi czy znamiona molestowania. Ale przede wszystkim uświadomił skalę problemu: 5 proc. spośród 200 tys. przeanalizowanych wiadomości zawierało nieodpowiednie sformułowania.

– Można bezpiecznie założyć, że jeden na dwadzieścia postów w mediach społecznościowych zawiera obraźliwe lub agresywne treści – mówi Gilles Jacobs.

**ŻEBY STWORZYĆ SYSTEM SPRAWNIE ANALIZUJĄCY JĘZYK,** nie wystarczy zebrać zdolnych programistów, analityków danych i ekspertów od przetwarzania języka. Do tego potrzebny jest interdyscyplinarny zespół składający się także z socjologów, psychologów czy kryminologów – ludzi rozumiejących mechanizmy cyberprzemocy. Prosta analiza danych może bowiem prowadzić do nadinterpretacji i przeregulowania. A jak

FOT.: PIOTR WANIOREK

tłumaczy Wroczyński, nie chodzi o stworzenie programu blokującego zwroty „k... mać" czy „ja p...", tylko nienawistne konstrukcje typu „szkoda, że cię matka nie wyabortowała".

Twórca Samurai Labs sztuczną inteligencją zajmuje się już ponad dwie dekady. Do branży zwalczania przemocy w sieci trafił nietypową drogą – poprzez czaty erotyczne. Opracował system, który miał konwersować z użytkownikami spragnionymi seksualnych doznań. Dość niespodziewanie okazał się on podwaliną pod jego późniejsze projekty, tworzone już w znacznie wznioślejszych celach. Najpierw we współpracy z Komendą Stołeczną Policji opracował system CERBER, służący do wychwytywania aktywności pedofilskiej na czatach. Potem uruchomił ISPAD, program do wykrywania agresji w internecie. Wreszcie przeniósł się do Kalifornii, gdzie powołał do życia Fido.ai – firmę pracującą nad sztuczną inteligencją potrafiącą wnikliwie analizować język internautów.

**5 PROC.**

**PUBLIKOWANYCH W INTERNECIE POSTÓW**
zawiera obraźliwe lub agresywne treści

Te wszystkie doświadczenia – razem z własnościami intelektualnymi – zebrał w Samurai Labs, spółce, w którą w 2018 roku zainwestowali przedsiębiorca Dawid Urban oraz piłkarz Robert Lewandowski. Postawił sobie za cel stworzenie systemu zwalczającego przemoc wobec dzieci: przeciwdziałającego m.in. szantażom, wymuszeniom czy nakłanianiu do samobójstw.

**WROCZYŃSKI ZAPRZĄGŁ DO WALKI TZW. TRZECIĄ FALĘ SZTUCZNEJ INTELIGENCJI** – wykorzystującą nie tylko deep learning (algorytmy samodzielnie uczące się na podstawie danych), ale także wnioskującą na podstawie opinii ekspertów.

– Nad podobnymi narzędziami pracowaliśmy wcześniej w ramach resortowych i naukowych projektów badawczo-rozwojowych. Teraz przygotowujemy rozwiązanie, które będzie można wdrożyć na masową skalę – mówi Michał Wroczyński.

Twórca Samurai Labs uważa, że jest w stanie skutecznie stawić czoła epidemii, ponieważ jego system miałby działać w czasie rzeczywistym – a więc przed wysłaniem krzywdzącej wiadomości, a nie post factum, jak w przypadku większości tego typu rozwiązań. Wierzy też, że może zbudować na nim wielki biznes. Wroczyński chce, aby Samurai Labs stał się tarczą antyprzemocową (wzorem programów antywirusowych), która byłaby wykorzystywana w aplikacjach, grach czy na stronach internetowych.

– Skoro osłaniamy nasze komputery, dlaczego mielibyśmy nie zapewnić ochrony dzieciom? – pyta Wroczyński. – Zaczynamy od przemocy wobec dzieci. Potem będziemy chronić też innych użytkowników internetu przed rozprzestrzeniającymi się formami przemocy, oszustwami, fake newsami i nienawiścią.

Przedsiębiorca uważa, że jego rozwiązanie może być skuteczniejsze od narzędzi stosowanych przez gigantów technologicznych, którzy często szukają metody niezależnej językowo (Samurai skupia się na angielskim). W ubiegłym roku przesłuchiwany przez senackie komisje Mark Zuckerberg stwierdził, że Facebook potrzebuje 5–10 lat, żeby zbudować narzędzie AI potrafiące wychwytywać w postach niuanse językowe. Dziś nad bezpieczeństwem użytkowników platformy czuwa ponad 30 tys. osób, których wspierają automatyczne narzędzia do wyławiania niepożądanych treści.

MARK ZUCKERBERG **zakłada, że Facebook potrzebuje 5-10 lat,** żeby opracować sztuczną inteligencję, która będzie sprawnie wychwytywać niuanse językowe i tym samym zapobiegać hejtowi na platformie

– W drugim kwartale 2018 roku usunęliśmy 2,5 mln wpisów zawierających mowę nienawiści. W trzecim kwartale 2018 r. liczba ta wzrosła już do 2,9 miliona – wylicza Jakub Turowski, dyrektor ds. polityki publicznej Facebooka w Polsce i krajach bałtyckich.

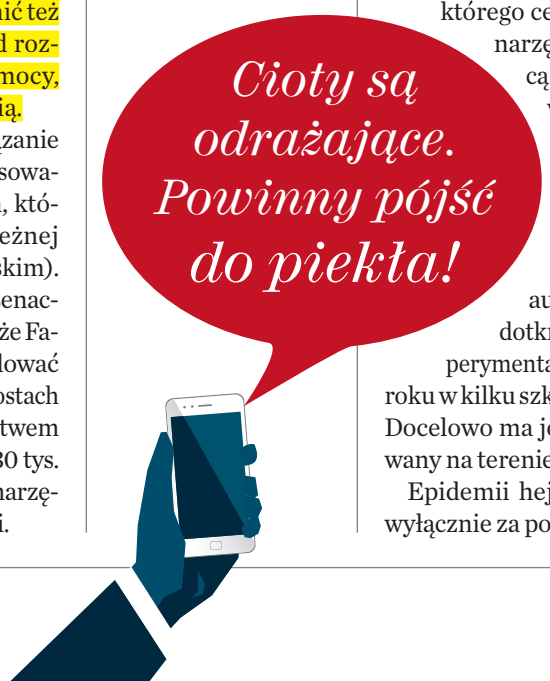
Na początku tego roku, dokonując czwartej oceny unijnego kodeksu postępowania przeciwko mowie nienawiści w sieci, Komisja Europejska pochwaliła tech-gigantów (Facebook, Microsoft, Twitter oraz YouTube) za walkę z problemem. Od 2016 roku na ich platformach odsetek naruszonych treści rozpatrywanych w ciągu 24 godzin wzrósł z 40 do 89 proc., a wskaźnik usuwania komunikatów nawołujących do nienawiści podniósł się z 28 do 72 procent. To pokrzepiające, choć dotyczy jedynie likwidacji szkód po pożarze, a nie zapobiegania podpaleniom.

Zresztą nawet całkowite oczyszczenie social mediów z hejtu nie rozwiąże problemu.

– Sieci społecznościowe mogą banować użytkowników, jeżeli ci nie przestrzegają reguł, ale na takich samych zasadach nie są w stanie funkcjonować firmowe intranety czy szkolne systemy. W takich przypadkach musimy pracować nad prewencją i edukacją – mówi Enrico Maria Parias, jeden z liderów projektu CREEP (Cyberbullying Effects Prevention).

**TO PANEUROPEJSKIE PRZEDSIĘWZIĘCIE** (zaangażowały się w nie m.in. włoska fundacja Bruno Kesslera, francuski instytut badawczy Inria czy niemiecki start-up NeuroNation), którego celem jest stworzenie zestawu narzędzi do walki z cyberprzemocą. Grupa naukowców – finansowanych przez unijną platformę EIT Digital – pracuje nad semantycznym systemem sztucznej inteligencji wychwytującym zagrożenia w sieci oraz chatbotem, który będzie automatycznie wspierał osoby dotknięte cyberprzemocą. Ten eksperymentalny projekt wystartował w 2018 roku w kilku szkołach we włoskim Trydencie. Docelowo ma jednak zostać zaimplementowany na terenie całej Europy.

Epidemii hejtu nie poskromi się jednak wyłącznie za pomocą technologii.




FOT.: J. SCOTT APPLEWHITE/AP PHOTO/EAST NEWS

Zdaniem prof. Michała Bilewicza, psychologa z Centrum Badań nad Uprzedzeniami UW, potrzebne są też znaczące zmiany w rzeczywistości politycznej i biznesowej.

– Mowa nienawiści musiałaby przestać się opłacać. A nie odnoszę wrażenia, żeby dostawcy informacji, właściciele portalów społecznościowych czy politycy rzeczywiście chcieli się pozbyć hejtu. Robią tylko tyle, ile muszą, aby nie podlegać penalizacji – mówi prof. Bilewicz.

Psycholog porównuje to do handlu krwawymi diamentami – wszyscy ludzie wiedzą, że kamienie ociekają krwią, że wydobywane są dzięki niewolniczej pracy Afrykanów, ale mimo to wielu nadal chce na nich zarabiać. Stąd biorą się agencje PR zatrudnione do znieważania ludzi i firm, przypominające ściek komentarze pod artykułami czy płatny trolling, który może kształtować nastroje społeczne, a nawet wpływać na wyniki wyborów.

*Matka zapomniała cię wyabortować!*
*Ty dzbanie!*

– Maszynowe mechanizmy rozumienia tekstu pomogłyby w walce z epidemią, gdyby taka była wola po stronie polityków i przedsiębiorców. Ale tej woli nie ma. Jedni widzą w hejcie zysk polityczny, a drudzy finansowy – mówi prof. Michał Bilewicz. I dodaje: – Dobrze, że pisze pan ten artykuł, ponieważ przedsiębiorcy powinni się zastanowić nad społecznymi konsekwencjami swoich działań. To koszt, którego żaden CSR nie pokryje. F

**KRZYSZTOF DOMARADZKI**