Pomorski Przegląd Gospodarczy

Sztuczna inteligencja dla bezpieczeństwa najmłodszych

MICHAŁ WROCZYŃSKI

Prezes Zarządu, Samurai Labs

GNIEWOSZ LELIWA

Dyrektor ds. Badań nad AI, Samurai Labs

Rękawicę w walce o bezpieczeństwo najmłodszych w internecie podjęło trójmiejskie laboratorium sztucznej inteligencji Samurai Labs. Ich autorski system przeciwdziałania przemocy w sieci funkcjonuje m.in. na największym forum internetowym Reddit. Jaki jest klucz do sukcesu w walce z cyberprzemocą? Na czym polega trzecia fala sztucznej inteligencji i czym różni się od poprzednich? W jaki sposób odnaleźć w sieci balans między bezpieczeństwem użytkowników a wolnością wypowiedzi?

Rozmowę prowadzi Marcin Wandałowski – redaktor prowadzący „Pomorskiego Przeglądu Gospodarczego”.

Prowadzona przez Panów firma – Samurai Labs – określana jest mianem „laboratorium sztucznej inteligencji”. Co się kryje za tym hasłem?

MW: Start-upy są organizacjami charakteryzującymi się bardzo dużą zwinnością i elastycznością. Dzięki tym cechom są one nieraz w stanie zagrozić biznesowo wielkim korporacjom – dzieje się to zresztą nie od dziś. Nowym trendem jest natomiast to, że w start-up’owym stylu otwiera się także laboratoria – miejsca, w których tworzy się naukę. W ten sposób takie start-up’owe laboratoria mogą konkurować na płaszczyźnie naukowej z wielkimi uniwersytetami czy działami badawczo-rozwojowymi potężnych korporacji.

GL: Tego typu laboratoria biorą to, co najlepsze ze świata akademickiego – m.in. metodę naukową, podejście do problemu, które przynosi obiektywne wyniki. Pozostają jednak przy tym podmiotami bardzo zwinnymi. Działające w ten sposób firmy nazywane są mianem *deep tech* start-upów.

MW: Powstają one po to, by rozwiązać konkretny, poważny problem – naukowy lub społeczny. Po osiągnięciu celu otwiera się pole dla wielu nowych, nieraz przełomowych szans biznesowych. Start-upy *deep tech* tworzą technologię umożliwiającą zrobienie czegoś, co wcześniej nie było możliwe – zajmują się m.in. opracowywaniem szczepionek, rozwijaniem pojazdów autonomicznych czy wirtualnej rzeczywistości.

Co jest przedmiotem badań laboratorium Samurai Labs?

MW: Doskonalimy narzędzie Samurai, potrafiące wykryć oraz zareagować w czasie rzeczywistym na internetową przemoc wobec dzieci, propozycje pedofilskie oraz – co niestety jest dziś w świecie online prawdziwą epidemią – zachowania samobójcze, takie jak deklaracje, plany, pytania o skuteczne sposoby i tym podobne. Opracowana przez nas technologia jest pionierską w skali świata i opiera się na tzw. trzeciej fali sztucznej inteligencji (*artificial intelligence* – AI).

Na czym polega trzecia fala AI i czym różni się od wcześniejszych?

GL: Amerykańska rządowa Agencja Zaawansowanych Projektów Badawczych w Obszarze Obronności – DARPA (*Defence Advanced Research Projects Agency*) wyróżnia trzy fale AI. Pierwsza, której największe sukcesy przypadają na lata 80. ubiegłego wieku, była oparta na systemach regułowych, gdzie całą wiedzę do systemu dostarczali ludzie – naukowcy i inżynierowie. Taka AI była w stanie wnioskować w oparciu o tę wiedzę, natomiast nie była w stanie sama się uczyć.

Później – aż do współczesności – rozwiązania AI opierały się głównie na uczeniu maszynowym. W tym podejściu system uczy się z danych oznaczonych przez ludzi, jak rozpoznawać pewne wzorce. Pokazując systemowi zdjęcia kota uczymy go, jak rozpoznawać te zwierzęta na nowych, niewidzianych wcześniej zdjęciach. Zaletą drugiej fali AI są bardzo duże możliwości uczenia się systemu, wady są jednak związane z tym, że system nie ma możliwości wnioskowania i abstrahowania. Jeśli w danych „uczących" nie pojawił się dany przypadek zachowania, system nie będzie w stanie sam go wykryć.

Trzecia fala, zwana też adaptacją kontekstualną (ang. *contextual adaptation*), w której działa Samurai Labs, skupia się na tym, by system mógł zarówno się uczyć (jak w drugiej fali), ale też i wnioskować w oparciu o wiedzę ekspertów (psychologów, lingwistów), będąc w stanie z dużą precyzją interpretować nowe, nieznane mu wcześniej przykłady (jak w pierwszej fali).

> Trzecia fala AI, zwana też adaptacją kontekstualną, skupia się na tym, by system mógł zarówno uczyć się z danych, ale też i wnioskować, będąc w stanie z dużą precyzją interpretować nowe, nieznane mu wcześniej przykłady.

Aby rozwijać rozwiązania AI trzeciej fali, potrzeba więc zatem przede wszystkim danych oraz wiedzy?

MW: Kluczowa jest wiedza – dlatego też zespół Samurai Labs jest wielodyscyplinarny. Na swoim pokładzie mamy matematyków, filozofów, specjalistów od sztucznej inteligencji, od rozumienia języka i przetwarzania języka naturalnego, psychologów, pedagogów. Dopiero takie połączenie, taki różnorodny zespół jest w stanie wszczepić w tworzone narzędzie wiedzę na temat tego, w jaki sposób są w internecie atakowane dzieci. Sposoby opisywania myśli samobójczych czy techniki postępowania pedofili, takie jak izolacja dziecka od opiekunów czy próba przekonania go, że jest już osobą dorosłą, są znane specjalistom z fachowej literatury i wieloletniego doświadczenia. Zbieramy tę wiedzę od ekspertów i „wkładamy" ją do maszyny. Dzięki niej oraz zdolności systemu do wnioskowania, potrzebujemy znacznie mniej danych. Nasz Samurai to *de facto* połączenie trzech światów – wiedzy eksperckiej, danych (choć w relatywnie niewielkiej ilości) oraz rozumienia języka.

> W Samurai Labs mamy matematyków, filozofów, specjalistów od AI, od rozumienia języka, od przetwarzania języka, psychologów, pedagogów. Dopiero taki różnorodny zespół jest w stanie wszczepić w tworzone narzędzie wiedzę na temat tego, w jaki sposób są w internecie atakowane dzieci.

Gdyby nasz pomysł opierał się na *machine learning* (drugiej fali AI) – gdzie absolutnie kluczowym elementem są dane, z których uczą się maszyny – nie bylibyśmy w stanie skutecznie go wdrożyć. Nikt bowiem nie dysponuje wystarczającą ilością danych przedstawiających sposoby komunikowania się pedofili czy technik szantażystów. Mamy wystarczająco dużo danych dotyczących korków na drogach, aby trenować system nawigacji, ale zjawiska, którymi my się zajmujemy, są relatywnie rzadsze (choć same w sobie są przerażające), więc dostęp do nich jest znacznie bardziej utrudniony.

W jaki sposób udaje się Wam konkurować z globalnymi potentatami technologicznymi, rozwijającymi podobnego typu rozwiązania?

MW: Źródłem naszej przewagi konkurencyjnej jest skuteczniejsza od konkurentów, wyższej jakości technologia. Giganci pokroju Google'a czy Facebooka, dysponując gigantycznymi zasobami danych, rozwijają narzędzia oparte na wspomnianym uczeniu maszynowym. Ich systemy uczą się, trenują na tych danych. Związane są z tym jednak dwa problemy. Po pierwsze – takie rozwiązania są mało skuteczne. Według różnych szacunków 50-70% wykrywanych przez nie przypadków to fałszywe alarmy – systemowi błędnie wydaje się, że treść jest obraźliwa, atakująca. Po drugie – jest to model działający w modelu *post factum*. Gdy pedofil próbuje zaatakować dziecko

przez sieć, ktoś musi to zobaczyć i zaraportować, a następnie AI podpowiada moderatorowi, czy dane sformułowanie uznajemy za niebezpieczne czy nie. Finalnie decyzję o usunięciu bądź pozostawieniu treści podejmuje moderator – tego rozwiązania nie da się zatem zastosować bez człowieka, a ten bywa niestety w świecie technologii tzw. wąskim gardłem: nie skaluje się, opóźnia pracę, dużo kosztuje. Co więcej, jest to bardzo trudna, niesamowicie obciążająca psychicznie praca, do wykonywania której nie ma wystarczająco dużo chętnych.

> W świecie technologii człowiek bywa często tzw. wąskim gardłem: nie skaluje się, opóźnia pracę, dużo kosztuje. Dlatego postawiliśmy na automatyzację.

Samurai jest znacznie bardziej skuteczny – dzięki włożonej w niego specjalistycznej wiedzy, na 100 badanych przypadków, tylko 4-6 zinterpretuje w błędny sposób. Co więcej – działa w sposób automatyczny jako „autonomiczny moderator". W praktyce oznacza to, że gdy ktoś kogoś będzie chciał zaatakować, nasz system zareaguje w czasie rzeczywistym i nie pozwoli wysłać danej wiadomości czy opublikować danego tekstu. System rozpozna również, kiedy osoba chcąca popełnić samobójstwo opublikuje swój pożegnalny list – stwarza to szansę na zmitygowanie sytuacji, na „podanie ręki" tej osobie, na połączenie jej z wyszkolonym w takich sytuacjach operatorem. Fizyczna interwencja moderatora następuje w bardzo nielicznych przypadkach, których narzędzie nie jest w stanie samodzielnie zinterpretować.

Czy Samurai jest już dostępny na rynku, czy trwają jeszcze prace nad jego udoskonalaniem?

MW: Mamy już kilka wdrożeń – pilotażowo uruchomiliśmy naszego Samurai'a na serwisie Reddit, który odwiedza dziennie kilkaset milionów użytkowników. W ramach projektu zleconego przez brytyjską policję pracujemy z Hatelabem, wykrywając internetową przemoc wobec Polaków i innych mniejszości, związaną z tematem brexitu.

Czy stronom internetowym pokroju Reddita zależy w ogóle na tym, by zwalczać przemoc na swoim serwisie? Często przecież znajdujemy w sieci formułki, głoszące że „redakcja nie ponosi odpowiedzialności za treści komentarzy" itp.

MW: Gdy w 2012 r., w ramach działalności wcześniejszego projektu – Fido Intelligence – tworzyliśmy pierwszy na świecie system do wykrywania aktywności pedofilów na czatach internetowych oraz przemocy, rynek nie był jeszcze na niego gotowy. Panowało wówczas przekonanie, że im więcej kłótni i obrażania, tym większy ruch na stronie i większa liczba kliknięć w reklamy.

Obecnie natomiast nasz *timing* jest idealny – ten, kto zarządza wspólnotą internetową chce, by była ona „zdrowa", pozbawiona przemocy i hejtu. Ma to wymiar pragmatyczny – według badań aż 20 proc. użytkowników zaatakowanych online na danej stronie internetowej czy aplikacji, rezygnuje z uczestnictwa w niej i odchodzi. Drugi argument jest natomiast związany z regulacjami prawnymi, szczególnie w Unii Europejskiej. W krajach takich jak Niemcy, Francja czy Wielka

Brytania, w myśl obowiązujących przepisów dużym wydawcom internetowym grożą gigantyczne, sięgające 50 mln euro kary za niereagowanie na przemoc online czy niezdejmowanie mowy nienawiści znajdującej się na ich łamach.

> Ten, kto zarządza wspólnotą internetową chce, by była ona „zdrowa", pozbawiona przemocy i hejtu. Ma to wymiar pragmatyczny – według badań aż 20 proc. użytkowników zaatakowanych online na danej stronie internetowej czy aplikacji, rezygnuje z uczestnictwa w niej i odchodzi.

W jakich językach działa Wasz system?

MW: W języku angielskim i częściowo polskim. Generalnie jednak skupiamy się na tym pierwszym, cały czas doskonalimy narzędzie w oparciu o niego, co swoją drogą również uważam za naszą przewagę konkurencyjną. Jako start-up dobrze jest być skoncentrowanym na pewnej konkretnej rzeczy, starając się być w niej najlepszym. Wierzę, że to dla nas dobra ścieżka. Choć oczywiście w miarę rozrostu firmy być może poszerzymy naszą ofertę o kolejne języki.

Czy nie boicie się, że Samurai mógłby wpaść w niepowołane ręce i zostać wykorzystany w złym celu – np. do cenzurowania wypowiedzi?

MW: Technologie niosą za sobą gigantyczne ryzyka nadużycia, jesteśmy tego zresztą świadkami – najlepszy przykład to systemy AI zajmujące się trollingiem, niszczeniem internetowych dyskusji. Wracając do Samuraia – „uzbrajając" go w odpowiednią wiedzę, mógłby służyć w różnych innych celach niż szukanie przemocy w sieci, np. w marketingu. Wyobrażam też więc sobie, że mógłby zostać użyty do celów niecnych. Nie udostępniamy jednak klientom narzędzia z opcją „zrób co chcesz" – sprzedajemy im bardzo konkretną usługę wirtualnego strażnika, który reaguje tylko wtedy, gdy jest świadkiem przemocy czy innych niepokojących zjawisk, o których mówiliśmy wcześniej.

Czy można powiedzieć, że reagowanie na przemoc w sieci – choć jest w założeniu szlachetną ideą – w praktyce przypomina balansowanie między bezpieczeństwem, a wolnością wypowiedzi?

MW: Zdecydowanie – bardzo nie chcielibyśmy, by Samurai był postrzegany jako narzędzie cenzurujące, ograniczające komunikację, lecz żeby reagował tylko wtedy, gdy jest to faktycznie uzasadnione. Nauczyliśmy się rozróżniać opinię – do której każdy ma prawo – od przemocy, którą staramy się wyeliminować. Mimo, że jest to subtelna różnica, nasz system w przeciwieństwie do większości innych, radzi sobie z nią bardzo dobrze. Jedyną receptą na to, by nie przekroczyć tej cienkiej granicy między bezpieczeństwem a wolnością, jest doskonalenie systemu tak, by popełniał jak najmniej błędów. Nam udaje się to jak na razie bardzo dobrze.

O rozmówcach

MICHAŁ WROCZYŃSKI

Prezes Zarządu, Samurai Labs

Michał Wroczyński – lekarz, terapeuta, kognitywista i futurolog. Od ponad 15 lat tworzy systemy sztucznej inteligencji oraz rozumienia języka naturalnego. Twórca czterech start-up'ów, w tym laboratorium sztucznej inteligencji Samurai Labs, które zajmuje się wykrywaniem i przeciwdziałaniem przemocy w internecie. Swój czas dzieli na dwa domy – w Trójmieście i w San Francisco.

GNIEWOSZ LELIWA

Dyrektor ds. Badań nad AI, Samurai Labs

Gniewosz Leliwa – dyrektor ds. badań nad sztuczną inteligencją i współzałożyciel Samurai Labs. W 2010 r. porzucił karierę akademicką w fizyce kwantowej, aby pracować nad sztuczną inteligencją i rozumieniem języka naturalnego. Autor pięciu patentów z dziedziny AI i NLP. Jego praca i zainteresowania badawcze koncentrują się wokół tzw. trzeciej fali sztucznej inteligencji według klasyfikacji DARPA.