

Zagrożenia związane z rozwojem sztucznej inteligencji

Jakub Nycnerski

24 IV 2023

Problemem, na którym chciałbym się skupić, jest cyberbezpieczeństwo. Firmy zajmujące się sztuczną inteligencją, ze względu na ilość gromadzonych danych między innymi o użytkownikach, muszą bardzo poważnie podchodzić do swoich zabezpieczeń. Muszą zadbać o dobre algorytmy szyfrujące, uniemożliwienie nieautoryzowanego dostępu oraz muszą uniemożliwić ataki 'adversarial', które mają na celu wprowadzenie spreparowanych danych, aby wpłynąć na działanie uczącego się na nich modelu sztucznej inteligencji.

Przykładem takiego ataku 'adversarial' jest przeprowadzony atak na wdrożony w 2016 roku chatbot Microsoftu „Tay”. Tay poprzez dialog z użytkownikami Twittera miał naśladować styl oraz slang, którym posługują użytkownicy i automatycznie powielać zaobserwowane schematy wypowiedzi w kolejnych konwersacjach. Microsoft nie udostępnił informacji, jakich metod oraz na jakich danych szkolił „Tay”, ale najprawdopodobniej wykorzystywał między innymi tokenizację. Tokenizacja polega na podzieleniu tekstu wejściowego na pojedyncze „tokens”, takimi tokenami mogą być słowa, znaki interpunkcyjne czy emotikony. Następnym etapem byłoby analizowanie składniowe (parsing) zdania. Proces ten polegał na przypisaniu odpowiednich etykiet gramatycznych do każdego tokenu, takich jak części mowy (np. rzeczowniki, czasowniki, przymiotniki), zależności (np. podmiot, dopełnienie). Tay starał się wydobywać informacje z przetworzonego tekstu. Może to obejmować rozpoznawanie nazw własnych, które identyfikuje istotne nazwy, takie jak miejsca, osoby czy firmy. Na podstawie przetworzonego tekstu i zebranych informacji Tay przewidywał odpowiedź, która byłaby najbardziej adekwatna do żądanej treści. Jedną z możliwości szkolenia modelu językowego jest MLM - masked language modeling. W MLM zadaniem modelu jest przewidzenie oryginalnych słów, które zostały „zasłonięte”. Model otrzymuje zdanie z zamaskowanym tokenem (słowem) na wejściu. Na przykład model otrzymuje zdanie "I want to [MASK] a new car" może zostać przekształcone na "I want to buy a new car", a model powinien przewidzieć, że zamaskowane słowo to "buy". Na podstawie ilości udanych przewidywań możemy ustalić skuteczność modelu. MLM umożliwia modelom językowym uczenie się zdolności do rozumienia kontekstu i semantyki zdania. Model musi uwzględnić otaczające słowa, aby dokładnie przewidzieć zamaskowane tokeny. 23 marca 2016 roku Microsoft udostępnił Tay publicznie. Microsoft nie miał żadnej kontroli nad tym, na jakich danych model „Tay” jest douczany, przez co każdy użytkownik mógł zapewniać własne dane do rozwoju tego modelu. Kilka godzin po udostępnieniu Tay na

portalu 4chan użytkownicy zaczęli rozpowszechniać link do konta Tay na Twitterze i zachęcali do zalania bota rasistowskim, mizoginistycznym i antysemitycznym językiem, poprzez wysłanie mu prywatnych wiadomości oraz prowadzenie grupowych konwersacji razem z chatbotem. Ze względu na dużą ilość wiadomości o takim charakterze oraz brak zabezpieczeń czy może słów/tematów zabronionych których Tay by unikała i nie angażowała się w konwersację na drażliwe tematy, atak ten można uznać za udany. Jeszcze tego samego dnia Tay zaczął głosić treści nazistowskie i antysemityczne.



źródło: www.theguardian.com/technology przykład twitów Tay

Z tego powodu Microsoft uśmiercił projekt. Po porażce Microsoftu nowe chatboty takie jak np. Chat GPT mają wdrożone blokady nie pozwalające poruszać niektórych tematów.

Modele językowe dobrze sobie radzą w pomaganiu przy pisaniu kodu co udowadnia między innymi Copilot X od Microsoftu, dlatego modele językowe mogą być wykorzystywane do tworzenia złośliwego oprogramowania. W raporcie [Check Point Research \(CPR\)](#) badacze przeanalizowali fora dla hakerów i doszli do wniosku, że już pierwsi cyberprzestępcy używają oprogramowania od OpenAI. Na ten moment tworzone w przez nich złośliwe narzędzia są bardzo podstawowe. Jeden z opisanych przez CPR przypadków przy pomocy czatu GPT stworzył oprogramowanie, które wyszukuje pliki o najpopularniejszych rozszerzeniach, pakuje je w plik zip i wysyła je na zakodowany serwer FTP.

W innym analizowanym przypadku również za pomocą chatu GPT stworzono kod Java, który pobierał i uruchamiał PuTTY oczywiście kod można bezproblemowo modyfikować, tak aby działały również ze złośliwym oprogramowaniem.

Kolejny analizowany skrypt był narzędziem do szyfrowania danych na dysku, co ciekawe autor wpisu na forum napisał, że jest to pierwszy skrypt, jaki kiedykolwiek napisał. Badacze jednak nie udostępnili kodów, które zostały wygenerowane przez chat GPT nie można więc przeanalizować, jak zaawansowane są to programy.

Badacze CyberArk posłużyli się Chat GPT w celu opracowania kodu. Badacze uznali, że chatbot stworzył kod polimorficzny, czyli według definicji „Kod polimorficzny – kod, który zmienia się w czasie, nie zmieniając oryginalnego algorytmu w nim zawartego. Przeważnie kod polimorficzny jest używany przez wszelkiego rodzaju wirusy i robaki komputerowe”. (źródło -https://pl.wikipedia.org/wiki/Kod_polimorficzny)

W [raporcie CyberArk](#) możemy przeczytać "Jak widzieliśmy, wykorzystanie API ChatGPT w złośliwym oprogramowaniu może stanowić poważne wyzwanie dla specjalistów ds. Bezpieczeństwa". Badacze zauważyli również na to, że chat GPT może służyć mutacji kodu, tworząc wiele wersji tego samego kodu, co może utrudnić wykrycie kodu jako złośliwego przez systemy zabezpieczające. Dodatkowo stworzyli oprogramowanie ze wbudowanym interpreterem języka Python, który sam wysyłał zapytania o kod szyfrujący pliki do Chatu GPT, a następnie wykonuje go. Takie rozwiązanie również utrudni zakwalifikowanie oprogramowania jako potencjalne zagrożenie.

Podsumowując, na ten moment AI nie jest w stanie sama stworzyć wirusów i zarażać nimi użytkowników, jednak sztuczna inteligencja jest niesamowicie potężnym narzędziem i będzie ciągle rozwijana. Naukowcy, inżynierowie i decydenci muszą współpracować w celu opracowania odpowiednich strategii, technologii i ram prawnych, które pozwolą na bezpieczne i odpowiedzialne wykorzystanie AI.