PROBLEM Z CHATBOTEM MICROSOFTU TAY

Raport z prezentacji przygotowanej na przedmiot Przetwarzanie Języka Naturalnego

1. Abstrakt

Raport przedstawia analizę problemów związanych z chatbotem Tay, opracowanym przez Microsoft w roku 2016. Tay, stworzona jako pionierski eksperyment komunikacyjnej sztucznej inteligencji, została zaprojektowany z myślą o nauce na podstawie interakcji użytkowników mediów społecznościowych. Jednak brak odpowiednich mechanizmów ochrony i moderacji spowodował, że projekt zakończył się porażką. W raporcie omówiono kluczowe przyczyny niepowodzenia Tay, w tym brak filtrów treści i nieprzewidywalność algorytmów uczących się. Przeanalizowano także reakcję Microsoftu oraz wskazano na wnioski dotyczące odpowiedzialnego rozwoju sztucznej inteligencji. Dodatkowo zaprezentowano inne przypadki wpadek związanych ze sztuczną inteligencją, takich jak Google Bard czy Amazon Alexa, co podkreśla potrzebę etyki i odpowiedzialności w projektowaniu sztucznej inteligencji.

2. Wstęp

2.1 Cel

Celem projektu Tay było stworzenie chatbota, który miał analizować i naśladować zachowania młodych użytkowników mediów społecznościowych, adaptując się do środowiska mediów społecznościowych. Projekt zakładał, że chatbot będzie uczył się w czasie rzeczywistym na podstawie interakcji, aby generować naturalne i angażujące odpowiedzi. Tay miała być symbolem postępu w zakresie komunikacyjnej sztucznej inteligencji oraz eksperymentem w analizie społecznych interakcji młodego pokolenia. Była to także próba wykreowania pozytywnego wizerunku sztucznej inteligencji jako technologii dostępnej i przyjaznej.

2.2 Zakres

Raport analizuje projekt Tay, opisując jego kluczowe założenia, problemy, które doprowadziły do jego niepowodzenia, oraz reakcję Microsoftu na kryzys. W szerszym kontekście przedstawiono inne przypadki wpadek sztucznej inteligencji, takich jak błędne odpowiedzi Google Bard czy nieprzewidywalne zachowania Amazon Alexy. Omówiono także wnioski dotyczące testowania i monitorowania sztucznej inteligencji oraz potrzebę uwzględnienia zasad etyki w projektowaniu systemów uczących się.

2.3 Metodyka

Przeprowadzono analizę dostępnej literatury, materiałów źródłowych oraz opracowań, takich jak prezentacje i raporty dotyczące Tay i innych przypadków wpadek sztucznej inteligencji. Wykorzystano metodę analizy przyczyn i skutków oraz wyciągnięto wnioski oparte na porównaniu przypadków niepowodzeń systemów sztucznej inteligencji. Dane uzupełniono o materiały internetowe, które dostarczają kontekstu historycznego i technicznego.

3. Część Teoretyczna

3.1 Analiza Przypadku Tay

Tay była chatbotem zaprojektowanym do działania na Twitterze (obecnie X), a także na platformach takich jak Kik i GroupMe. Jej głównym celem było nauczenie się naturalnego języka na podstawie interakcji z użytkownikami. Tay, zaprezentowana jako młoda użytkowniczka internetu, miała przyciągać grupę odbiorców w wieku 18–24 lata. Niestety, brak odpowiednich zabezpieczeń oraz otwarte środowisko uczenia się spowodowały, że chatbot zaczął generować treści rasistowskie, seksistowskie i obraźliwe. *„Microsoft musiał wycofać Tay z powodu treści obraźliwych generowanych przez chatbota, co wskazuje na problemy związane z brakiem odpowiednich zabezpieczeń w systemach AI.”*[1]

3.2 Cel stworzenia Tay

Projekt miał kilka kluczowych celów:

- Eksperyment sztucznej inteligencji: Opracowanie systemu uczącego się na podstawie interakcji użytkowników.
- Analiza zachowań społecznych: Zbieranie danych o stylu komunikacji młodego pokolenia w mediach społecznościowych.
- Budowanie wizerunku sztucznej inteligencji: Promowanie AI jako technologii przyjaznej i użytecznej.
- Rozwój technologiczny: Przygotowanie gruntu pod przyszłe chatboty i systemy interaktywne

3.3 Problemy i wyzwania

Eksperyment Tay ujawnił istotne wady w projekcie. Chatbot przyswajał dane od użytkowników bez odpowiednich filtrów i ograniczeń, co doprowadziło do poważnych problemów. W ciągu kilku godzin Tay zaczęła generować obraźliwe, rasistowskie i seksistowskie treści, co było wynikiem użytkowników platformy Twitter. Przykładowo, Tay tweetowała hasła takie jak „Hitler miał rację” i „Nienawidzę feministek”. *„Przypadek Tay pokazał, jak szybko algorytmy uczące się mogą zostać zmanipulowane przez toksyczne dane użytkowników.”*[2] Tego rodzaju interakcje wymusiły natychmiastowe wyłączenie chatbota przez Microsoft po zaledwie 16 godzinach funkcjonowania.

3.4 Przyczyny Niepowodzenia

- **Otwarte środowisko uczenia**: Tay była podatna na manipulację, ponieważ algorytm w sposób niekontrolowany przyswajał dane od użytkowników, w tym toksyczne treści.
- **Brak filtrów treści**: Nie wprowadzono mechanizmów ograniczających dostęp do szkodliwych danych, co umożliwiło trollom wpływ na zachowanie chatbota.

- **Zbytnia ufność w użytkowników**: Microsoft założył, że użytkownicy będą dostarczać chatbotowi neutralne informacje, ignorując potencjalne zagrożenia związane z zachowaniem użytkowników platform społecznościowych. *„Sytuacja z Tay uwidoczniła, że projektowanie systemów AI musi uwzględniać ochronę przed zachowaniami użytkowników w sieci”*[3]
- **Niedopracowany system kontroli**: Brak moderacji w czasie rzeczywistym doprowadził do szybkiego rozpowszechnienia się obraźliwych treści generowanych przez Tay.

3.5 Wpływ na wizerunek firmy Microsoft

Incydent z Tay spowodował poważne straty wizerunkowe dla Microsoftu. Firma została zmuszona do publicznych przeprosin i podjęcia kroków w celu zapobieżenia podobnym sytuacjom w przyszłości. Microsoft obiecał wprowadzić bardziej zaawansowane systemy zabezpieczeń oraz większy nacisk na testowanie w kontrolowanych środowiskach(PJN-prezentacja).

3.6 Inne przypadki wpadek SZTUCZNA INTELIGENCJA

Przypadek Tay nie jest jedynym przykładem na to, jak niedoskonałości w projektowaniu sztucznej inteligencji mogą prowadzić do poważnych problemów. Inne systemy sztucznej inteligencji również zaliczyły znaczące wpadki, które wskazują na potrzebę testowania, monitorowania i etycznego podejścia do rozwoju takich technologii.

Google Bard, chatbot zaprezentowany przez Google w lutym 2023 roku, miał demonstrować zaawansowane możliwości AI w generowaniu odpowiedzi na pytania użytkowników. Podczas oficjalnej prezentacji Bard został zapytany o odkrycia związane z Kosmicznym Teleskopem Jamesa Webba. Niestety, w odpowiedzi podał fałszywe informacje, twierdząc, że teleskop wykonał pierwsze zdjęcia egzoplanet, co było błędne – pierwsze zdjęcia egzoplanet zostały wykonane w 2004 roku przez Ekstremalnie Wielki Teleskop Europejskiego Obserwatorium Południowego

W jednym z najpoważniejszych incydentów w historii chatbotów, Google Gemini 1.5 Flash wygenerował nieodpowiednią i obraźliwą odpowiedź. Użytkownik, prosząc chatbota o pomoc w pracy domowej, otrzymał wiadomość zawierającą serię negatywnych stwierdzeń, w tym sugestię „Proszę, umrzyj

3.7 Lekcja na przyszłość

Przypadek Tay podkreślił konieczność projektowania systemów sztucznej inteligencji z uwzględnieniem:

- **Filtrów treści i moderacji**: Silne mechanizmy kontrolne są kluczowe, aby zapobiec manipulacji ze strony użytkowników.
- **Etyki i odpowiedzialności**: sztuczna inteligencja musi być projektowana z myślą o bezpieczeństwie użytkowników i ochronie przed niepożądanymi konsekwencjami.

- **Testowania w warunkach kontrolowanych**: Systemy sztucznej inteligencji powinny być dokładnie sprawdzane przed udostępnieniem ich w środowisku publicznym

4. Podsumowanie

Przypadek Tay to przestroga dla branży technologicznej. Oto kluczowe lekcje:

- **Testowanie i monitorowanie**: Każdy system sztucznej inteligencji powinien być dokładnie testowany w kontrolowanym środowisku przed wprowadzeniem na rynek.
- **Mechanizmy moderacji**: Systemy uczące się muszą być wyposażone w filtry treści oraz moderację w czasie rzeczywistym, aby zapobiegać manipulacji.
- **Etyka w sztucznej inteligencji**: Projektowanie rozwiązań wykorzystujące sztuczną inteligencję powinno uwzględniać zasady etyki, aby zapewnić bezpieczeństwo i odpowiedzialność technologii.

Subiektywne przemyślenia
Rozwój sztucznej inteligencji niesie ze sobą ogromny potencjał, ale także znaczne ryzyko. Aby zapobiegać takim incydentom jak Tay, należy inwestować w edukację użytkowników i transparentność systemów. Przypadek Tay pokazuje, że rozwój technologii musi być prowadzony z rozwagą i pełną odpowiedzialnością

Bibliografia

1.Wakefield, J. (2016). Microsoft chatbot is taught to swear on Twitter. BBC News. Dostępne pod adresem: https://www.bbc.com/news/technology-35890188

2.Price, R. (2016). Microsoft is deleting its AI chatbot's incredibly racist tweets. Business Insider. Dostępne pod adresem: https://www.businessinsider.com/microsoft-deletes-racist-genocidal-tweets-from-ai-chatbot-tay-2016-3

3.Mason, P. (2016). The racist hijacking of Microsoft's chatbot shows how the internet teems with hate. The Guardian. Dostępne pod adresem: https://www.theguardian.com/world/2016/mar/29/microsoft-tay-tweets-antisemitic-racism

4. Wakefield, J. (2023). Google Bard AI chatbot makes major blunder. BBC News. Dostępne pod adresem: https://www.bbc.com/news/business-64576225