Streszczenie popularnonaukowe

Projekt bada związki między językiem, uprzedzeniami rasowymi i sztucznej inteligencji (AI) w Polsce, ze szczególnym uwzględnieniem języka polskiego i jego użytkowników/użytkowniczek. Jego istotą jest zrozumienie, jak systemy AI, takie jak duże modele językowe (LLM) w języku polskim, mogą rozpowszechniać rasistowskie stereotypy i uprzedzenia wobec osób niebiałych. Celem projektu jest zwrócenie uwagi na to, jak uprzedzenia rasowe są wbudowane w językowe i komunikacyjne sposoby kształtowania rozumienia rasy w Polsce. Głównym celem projektu jest zrozumienie tego, co się dzieje wtedy, gdy algorytmy przetwarzania języka polskiego utrwalają rasistowskie stereotypy i uprzedzenia rasowe wobec osób niebiałych. Podstawą jest tutaj zbadanie złożonych *multimodalnych dyskursów rasizujących*, które odnoszą się do wielopoziomowych semantycznych i pragmatycznych procesów ideologicznych, w których ramach rasa jest konstruowana przez praktyki językowe, dyskursywne i komunikacyjne.

Przez analizę treści generowanych przez AI w języku polskim i zbadanie perspektywy różnorodnych etnicznie użytkowników/użytkowniczek języka polskiego, projekt podkreśla potrzebę bardziej sprawiedliwych i transparentnych systemów AI w przetwarzaniu języka polskiego. Przedmiotem badań są duże modele językowe (LLMs) takie jak polskojęzyczna wersja czatu GPT oraz wchodzący na rynek wytrenowany na polskojęzycznych danych PLLuM. Celem badań jest zidentyfikowanie uprzedzeń rasowych w treściach generowanych przez AI i zbadanie tego, jak te uprzedzenia wpływają na ludzi, szczególnie osób niebiałych. W świetle tych zagadnień, projekt ma na celu zbadanie związku między treściami generowanymi przez AI a *multimodalnymi dyskursami urasawiającymi*, które mogą utrwalać uprzedzenia rasowe w polskich dużych modelach językowych.

Główne cele projektu to:

(1) Zrozumienie, sproblematyzowanie i skonceptualizowanie tego, jak *multimodalne dyskursy urasawiające,* któr utrwalają uprzedzenia rasowe, są organizowane i reprezentowane w treściach generowanych przez AI, ze szczególnym uwzględnieniem czatu GPT i modelu PLLuM;
(2) Zbadanie tego, jak te paradygmaty są postrzegane i doświadczane przez różnorodnych etnicznie użytkowników/użytkowniczki języka polskiego;
(3) Zbadanie tego, jak modele językowe AI odzwierciedlają i kształtują uprzedzenia rasowe oraz jak proces ten przyczynia się to do szerszego zrozumienia społecznych implikacji technologii AI w kontekście polskim. Wyniki badań dostarczą ważnych informacji o tym, jak technologie te mogą utrwalać lub kwestionować istniejące stereotypy i nierówności;
(4) Zbadanie tego, jak *multimodalne dyskursy urasawiające* są obecne w polskojęzycznych treściach generowanych modeli sztucznej AI i jak są one doświadczane przez użytkowników/użytkowniczki, zwłaszcza tych ze społeczności niebiałych;
(5) Zrozumienie tego, jak polscy eksperci/ekspertki od sztucznej inteligencji i przetwarzania języka naturalnego postrzegają i zarządzają tymi uprzedzeniami podczas opracowywania takich systemów.

Większość badań w tej dziedzinie skupia się na kontekstach anglojęzycznych, często pomijając to, jak AI zachowuje się w językach o innej gramatyce i specyficznym kontekście kulturowym. Projekt te wypełnia tę lukę, badając uprzedzenia rasowe w tekstach i treściach wizualnych generowanych przez polską sztuczną inteligencję, jednocześnie próbując sprawić, by systemy te stały się bardziej inkluzywne. Projekt wykorzystuje mieszane metody, łącząc narzędzia lingwistyczne, socjologiczne i obliczeniowe do analizy nie tylko tekstu, ale także obrazów generowanych przez AI, dając pełniejszy obraz tego, jak uprzedzenia działają w treściach multimodalnych. Projekt umożliwi lepsze zrozumienie związków między technologią AI a polską tkanką uprzedzeń społecznych. Wyniki badań przyczynią się także do głębszego zrozumienia mechanizmów uprzedzeniowych, które działają w treściach generowanych przez polskojęzyczne duże modele językowe, a także umożliwią poznanie perspektyw i doświadczeń użytkowników/użytkowniczek języka. A to może wpłynąć na bardziej sprawiedliwe praktyki użycia AI w polskiej komunikacji. Celem projektu jest także stworzenie rozszerzenie wiedzy praktycznej w obszarze tworzenia polityk, oprogramowań i edukacji, aby promować bardziej inkluzywne środowiska cyfrowego oraz pogłębiać świadomości i dyskusji na temat odpowiedzialności etycznej związanej z rozwojem i wdrażaniem nowoczesnych technologii.

Nadrzędnym celem projektu jest ustanowienie nowych standardów badania uprzedzeń rasowych w nieanglojęzycznych systemach AI oraz dostarczenie metod i perspektyw, które mogą wyznaczać rozwój bardziej inkluzywnych technologii AI. Projekt bada również, jak technologie te mogą być wykorzystywane do zwalczania uprzedzeń w polskich modelach językowych, wyznaczając drogę do bardziej sprawiedliwej cyfrowej przyszłości w tym kontekście geograficznym.