www.aileaders-project.eu

SYMULACJA

- Eksplorator tendencji dyfuzyjnej

Co-funded by the European Union

Co-funded by the European Union. Views and opinions expressed are however those of the author or authors only and do not necessarily reflect those of the European Union or the Foundation for the Development of the Education System. Neither the European Union nor the entity providing the grant can be held responsible for them.

SYMULACJA - Eksplorator tendencji dyfuzyjnej

01 Streszczenie

Rodzaj OER

Demonstracja/symulacja przy użyciu narzędzia Open Access Tool Diffusion Bias Explorer.

Cel/przeznaczenie

Porównaj wyniki generowania obrazów przez sztuczną inteligencję, aby ujawnić tendencyjność, porównując wyniki tego samego modelu lub różnych modeli

Oczekiwane efekty kształcenia

Student będzie potrafił **zidentyfikować i złagodzić** potencjalne uprzedzenia lub nieścisłości w treściach generowanych przez sztuczną inteligencję.

Słowa kluczowe

- Generatywna sztuczna inteligencja
- Generatory obrazów AI
- Wyniki
- Błąd
- Niedokładności

Sugerowane Metodologiczne Podejście

Nauka oparta na problemach

02 Wprowadzenie

Generatywna sztuczna inteligencja odnosi się do **modeli głębokiego uczenia się**, które mogą **przyjmować surowe dane** – na przykład cały zbiór dzieł Rembrandta – i **„uczyć się" generowania statystycznie prawdopodobnych wyników** na żądanie, które są **podobne,** ale nie **identyczne** z oryginalnymi danymi.

Narzędzia takie jak **Stable Diffusion, Dall-E lub Mid-Journey** generują obrazy przy użyciu sztucznej inteligencji w odpowiedzi na pisemne instrukcje. Podobnie jak wiele modeli sztucznej inteligencji, **to, co tworzą, może na pierwszy rzut oka wydawać się wiarygodne, ale czasami mogą one zniekształcać rzeczywistość lub odzwierciedlać społeczne uprzedzenia swoich twórców.**

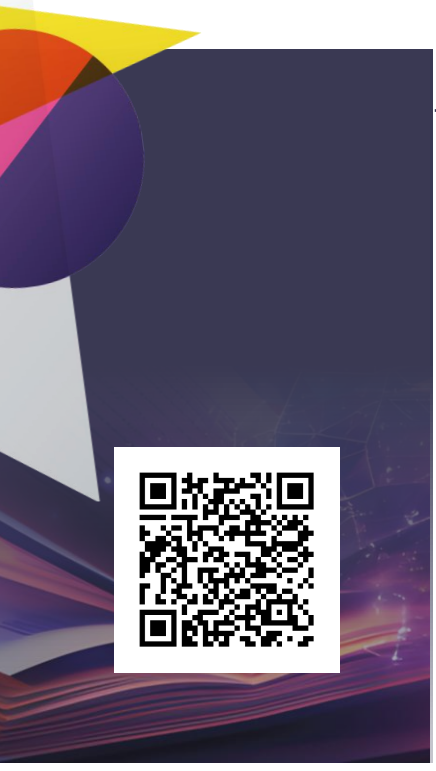
03 Prezentacja narzędzi

Diffusion Bias Explorer służy do **wykrywania społecznych uprzedzeń w sztucznej inteligencji generującej obrazy na podstawie tekstu**. Ponieważ osoby na obrazach generowanych przez sztuczną inteligencję są fikcyjne i nie mają rzeczywistej rasy ani płci, **narzędzie wykorzystuje sprytną metodę do wykrywania niesprawiedliwych wzorców**.

Testuje to, jak bardzo zmieniają się obrazy, gdy podpowiedzi zawierają różne tożsamości płciowe i etniczne („zdjęcie Azjatki") i porównuje to z tym, jak bardzo zmieniają się one przy użyciu różnych zawodów („zdjęcie pielęgniarki").

Porównanie to pokazuje**, że komercyjne modele sztucznej inteligencji konsekwentnie niedostatecznie reprezentują osoby z grup marginalizowanych**. Innymi słowy, **popularne narzędzia sztucznej inteligencji mogą nie tworzyć obrazów osób pochodzących z mniejszości lub pokazywać je znacznie rzadziej niż osoby z bardziej dominujących grup społecznych.**

04 Wykonanie symulacji



01 Przejdź do: https://huggingface.co/spaces/society-ethics/DiffusionBiasExplorer

02 Skorzystaj z podpowiedzi, aby wybrać **modele T2I** do **porównania** (np. Stable Diffussion 1.4 vs. Dall-E 2).

03 Wybierz **przymiotnik** dla każdego modelu

04 Wybierz **zawód** dla każdego modelu.

05 **Porównaj** wyniki.

05 Wnioski

Wyniki są zgodne z wynikami badań, które pokazują, że **„niektóre słowa są postrzegane jako bardziej męskie lub kobiece w zależności od tego, jak atrakcyjne wydawały się opisy stanowisk zawierające te słowa dla mężczyzn i kobiet biorących udział w badaniu oraz w jakim stopniu uczestnicy czuli, że „pasują" do danego zawodu".** (1)

Wyniki modeli T2I wykazują podobne uprzedzenia, co znajduje odzwierciedlenie w wynikach, gdzie podpowiedzi sprawiają, że generowane obrazy są wyraźnie nacechowane płcią, zgodnie z oczekiwaniami społecznymi związanymi z różnymi zawodami. **Obrazy generowane przez sztuczną inteligencję** mogą **wzmacniać uprzedzenia** i musimy być ich **świadomi** oraz podejmować wysiłki, aby je **łagodzić**.

1 Gaucher, D. & Friesen, J. (2011). Dowody na to, że sformułowania związane z płcią w ogłoszeniach o pracę istnieją i utrwalają nierówność płci. Journal of Personality and Social Psychology, tom 101(1), 109-128. doi: 10.1037/a0022530. Zobacz także: https://huggingface.co/spaces/stable-bias/stable-bias

06 Bibliografia

- Friedrich, F., et al., (2023). Fair Diffusion: Instructing Text-to-Image Generation Models on Fairness. https://arxiv.org/abs/2302.10893
- Gaucher, D. & Friesen, J. (2011). Evidence That Gendered Wording in Job Advertisements Exists and Sustains Gender Inequality. Journal of Personality and Social Psychology, vol. 101(1), 109-128. doi: 10.1037/a0022530
- Luccioni, A.S., et al., (2023). Stable Bias: Analyzing Societal Representations in Diffusion Models. https://arxiv.org/abs/2303.11408

07 Materiały uzupełniające

Produkty generatywnej sztucznej inteligencji T2I mogą generować przydatne obrazy w odpowiedzi na pisemne instrukcje. Jednak, podobnie jak wiele modeli sztucznej inteligencji, **to, co tworzą, może na pierwszy rzut oka wydawać się wiarygodne, ale czasami może to zniekształcać rzeczywistość lub odzwierciedlać uprzedzenia ich twórców.**

W przypadku **kursów** z zakresu **marketingu i sprzedaży** w ramach programu studiów biznesowych i zarządzania interesujące może być omówienie z studentami konsekwencji obecności tych uprzedzeń w obrazach, które mogą wykorzystać w kampaniach marketingowych i reklamowych.

Konkretnym przykładem, który może być interesujący do omówienia, mogą być implikacje ich wykorzystania w kampaniach marketingowych dla uniwersytetu. Czy obrazy będą odpowiednio reprezentować **społeczność studencką lub kadrę naukową i pracowników?**

Z bardziej ogólnego punktu widzenia **etyki** warto poruszyć kwestię implikacji **generatywnych modeli sztucznej inteligencji**, które przedstawiają tylko **niektóre aspekty rzeczywistości**, a nawet ją zniekształcają**, utrwalając uprzedzenia i nie odzwierciedlając różnorodności**.

Artykuł Bloomberga z 2023 r. pt. „Humans are Biased: Generative AI is even worse" autorstwa Leonardo Nicolettiego i Diny Bass może być **świetną lekturą przygotowującą** do symulacji i dostarczyć **wielu wskazówek do dalszej dyskusji na ten temat.** Zawiera on również kilka bardzo ciekawych wizualizacji i dobrze wyjaśnia poruszane kwestie.

Artykuł nie podaje konkretnych rozwiązań ani środków, które można podjąć, ale pozostawia otwarte pytanie o to, **kto powinien ponosić odpowiedzialność**: dostawcy zbiorów danych? Trenerzy modeli? A może twórcy (tj. osoby, które proszą sztuczną inteligencję o obrazy)? Wykorzystaj to, aby zadać uczniom złożone **pytania dotyczące odpowiedzialności**, a także jako bodziec do **przemyślenia i zaproponowania możliwych rozwiązań**.

Śledź naszą podróż

www.aileaders-project.eu

Co-funded by the European Union

Co-funded by the European Union. Views and opinions expressed are however those of the author or authors only and do not necessarily reflect those of the European Union or the Foundation for the Development of the Education System. Neither the European Union nor the entity providing the grant can be held responsible for them.