A.I.-tool voor het filteren van haat en discriminatie

Een A.I.-filter analyseert reacties op Instagram en waarschuwt gebruikers bij negatieve of discriminerende taal. Herhaalde overtredingen kunnen leiden tot een melding bij Instagram en mogelijke accountblokkade.

Created by: Tendys
Created on: March 19, 2025 12:02 PM
Changed on: March 31, 2025 4:54 AM

Context of use: Education
Level of education: Bachelor

Impact on society

What impact is expected from your technology?

This category is only partial filled.

What is exactly the problem? Is it really a problem? Are you sure?

Het probleem is dat kleinerende en discriminerende reacties gemakkelijk op Instagram geplaatst kunnen worden. Dit kan leiden tot symptomen zoals depressie, stress en angst, en draagt bij aan een toxische online omgeving. Onderzoek toont aan dat online negativiteit de mentale gezondheid schaadt en de gebruikerservaring verslechtert, wat bewijst dat dit een serieus probleem is.

Are you sure that this technology is solving the RIGHT problem?

This question has not been answered yet.

How is this technology going to solve the problem?

This question has not been answered yet.

What negative effects do you expect from this technology?

This question has not been answered yet.

In what way is this technology contributing to a world you want to live in?

This question has not been answered yet.

Now that you have thought hard about the impact of this technology on society (by filling out the questions above), what improvements would you like to make to the technology? List them below.

In de eerste testfase van het AI-filter kunnen sommige mensen onterecht worden uitgesloten en zal de app niet meteen alle negatieve reacties correct herkennen en filteren. Om dit te verbeteren, zou ik een breder getraind AI-model willen gebruiken met diverse testgroepen, zodat het filter inclusiever en accurater wordt.

Hateful and criminal actors

What can bad actors do with your technology?

This category is only partial filled.

In which way can the technology be used to break the law or avoid the consequences of breaking the law?

Als de A.I.-filter gehackt wordt, kan dit privacywetten schenden door mee te kijken met gebruikers of hun berichten te manipuleren. Daarnaast kan de filter worden omzeild, waardoor discriminerende reacties geplaatst kunnen worden zonder consequenties.

Can fakers, thieves or scammers abuse the technology?

This question has not been answered yet.

Can the technology be used against certain (ethnic) groups or (social) classes?

This question has not been answered yet.

In which way can bad actors use this technology to pit certain groups against each other? These groups can be, but are not constrained to, ethnic, social, political or religious groups.

This question has not been answered yet.

How could bad actors use this technology to subvert or attack the truth?

This question has not been answered yet.

Now that you have thought hard about how bad actors can impact this technology, what improvements would you like to make? List them below.

Ja, wij denken dat een goede oplossing zou zijn om de gebruiker de mogelijkheid te geven de AI-filter zelf in of uit te schakelen. Dit zou echter wel beveiligd moeten zijn met een wachtwoord om een veilige omgeving te waarborgen en te voorkomen dat kwaadwillende er zomaar mee aan de haal kunnen gaan.

Privacy

Are you considering the privacy & personal data of the users of your technology?

This category is only partial filled.

Does the technology register personal data? If yes, what personal data?

De A.I.-filtertool zelf gebruikt geen persoonlijke data. De enige gegevens die verwerkt worden, zijn de data die Instagram al verzamelt bij het installeren van de app.

Do you think the technology invades the privacy of the stakeholders? If yes, in what way?

This question has not been answered yet.

Is the technology is compliant with prevailing privacy and data protection law? Can you indicate why?

This question has not been answered yet.

Does the technology mitigate privacy and data protection risks/ concerns (privacy by design)? Please indicate how.

This question has not been answered yet.

In which way can you imagine a future impact of the collection of personal data?

This question has not been answered yet.

Now that you have thought hard about privacy and data protection, what improvements would you like to make? List them below.

Wij denken dat het een verbetering zou zijn als gebruikers van Instagram regelmatig en op een duidelijke manier op de hoogte worden gehouden van het privacy- en databeleid. Dit kan door middel van korte, begrijpelijke meldingen binnen de app, omdat veel gebruikers anders de voorwaarden niet lezen. Daarnaast zou een compacte en simpele uitleg van de AI-filter en hoe deze omgaat met gegevens helpen om transparantie te vergroten.

Human values

How does the technology affect your human values?

This category is only partial filled.

How is the identity of the (intended) users affected by the technology?

De A.I.-filtertechnologie zorgt ervoor dat gebruikers zich veiliger voelen op het platform, omdat kleinerende en discriminerende reacties worden tegengehouden. Dit kan ertoe leiden dat mensen zich vrijer uiten en hun identiteit zonder angst delen.

How does the technology influence the users' autonomy?

This question has not been answered yet.

What is the effect of the technology on the health and/or well-being of users?

This question has not been answered yet.

Now that you have thought hard about the impact of your technology on human values, what improvements would you like to make to the technology? List them below.

Wij denken dat de AI-filter al bijdraagt aan het verbeteren van online platformen door stress en angst te verminderen. Om dit verder te verbeteren, zouden we gebruikersfeedback verzamelen via korte formulieren op Instagram. Op basis van deze feedback kunnen we de AI-filter verfijnen, zodat deze nog beter aansluit bij de behoeften en waarden van de gebruikers.

Stakeholders

Have you considered all stakeholders?

This category is only partial filled.

Who are the main users/targetgroups/stakeholders for this technology? Think about the intended context by answering these questions.

Name of the stakeholder
Instagram gebruikers

How is this stakeholder affected?
-

Did you consult the stakeholder?
No

Are you going to take this stakeholder into account?
No

Did you consider all stakeholders, even the ones that might not be a user or target group, but still might be of interest?

-

Now that you have thought hard about all stakeholders, what improvements would you like to make? List them below.

Wij denken dat deze vraag niet direct op ons van toepassing is, omdat de AI-filter door alle Instagram-gebruikers gebruikt kan worden. Wel zouden we de filter verder kunnen verbeteren door meer maatwerkopties toe te voegen, zodat verschillende groepen gebruikers (zoals contentmakers, bedrijven en gewone gebruikers) hun instellingen beter kunnen afstemmen op hun behoeften.

Data

Is data in your technology properly used?

This category is only partial filled.

Are you familiar with the fundamental shortcomings and pitfalls of data and do you take this sufficiently into account in the technology?

Ik ben me ervan bewust dat A.I.-systemen misbruikt kunnen worden en dat er risicos zoals bias en privacy problemen kunnen ontstaan. Om deze tekortkomingen te beperken, zou ik een geavanceerde A.I. implementeren die context beter begrijpt en regelmatig wordt geüpdatet om nieuwe problemen te voorkomen. Daarnaast zou ik ervoor zorgen dat het systeem transparant werkt en gebruikers controle geeft over hoe het hun gegevens verwerkt.

How does the technology organize continuous improvement when it comes to the use of data?

This question has not been answered yet.

How will the technology keep the insights that it identifies with data sustainable over time?

This question has not been answered yet.

In what way do you consider the fact that data is collected from the users?

This question has not been answered yet.

Now that you have thought hard about the impact of data on this technology, what improvements would you like to make? List them below.

hier hebben we nog niet over na gedacht.

Inclusivity

Is your technology fair for everyone?

This category is only partial filled.

Will everyone have access to the technology?

This question has not been answered yet.

Does this technology have a built-in bias?

Het AI-filter van Instagram is bedoeld om voor alle gebruikers een veilige omgeving te creëren en wordt niet specifiek tegen één groep ingezet. Toch kan er in het begin sprake zijn van ingebouwde bias, omdat het AI-filter getraind wordt op data die mogelijk niet volledig neutraal is. Door gebruikersmeldingen en regelmatige updates kan het AI-filter bijgestuurd worden om eerlijker en nauwkeuriger te beoordelen

Does this technology make automatic decisions and how do you account for them?

This question has not been answered yet.

Is everyone benefitting from the technology or only a a small group? Do you see this as a problem? Why/why not?

This question has not been answered yet.

Does the team that creates the technology represent the diversity of our society?

This question has not been answered yet.

Now that you have thought hard about the inclusivity of the technology, what improvements would you like to make? List them below.

Wij denken dat de AI-filter al breed toegankelijk is voor alle Instagram-gebruikers. Toch zouden we de inclusiviteit kunnen verbeteren door ervoor te zorgen dat de AI beter werkt voor verschillende talen en culturen, en dat mensen met een visuele of cognitieve beperking de instellingen gemakkelijker kunnen gebruiken.

Transparency

Are you transparent about how your technology works?

This category is only partial filled.

Is it explained to the users/stakeholders how the technology works and how the business model works?

Het AI-filter is ingebouwd in Instagram en monitort reacties automatisch. Het waarschuwt bij ongepaste berichten en kan bij herhaaldelijk overtreden accounts tijdelijk verbannen. Instagram informeert gebruikers hierover via meldingen en community richtlijnen.

If the technology makes an (algorithmic) decision, is it explained to the users/stakeholders how the decision was reached?

This question has not been answered yet.

Is it possible to file a complaint or ask questions/get answers about this technology?

This question has not been answered yet.

Is the technology (company) clear about possible negative consequences or shortcomings of the technology?

This question has not been answered yet.

Now that you have thought hard about the transparency of this technology, what improvements would you like to make? List them below.

Als we terugkijken naar de eerder genoemde stappen, zien we dat de AI-filter al redelijk transparant is en dat we op dit moment geen verbeteringen kunnen bedenken.

Sustainability

Is your technology environmentally sustainable?

This category is only partial filled.

In what way is the direct and indirect energy use of this technology taken into account?

Dit sluit niet aan bij mij ai model

Do you think alternative materials could have been considered in the technology?

This question has not been answered yet.

Do you think the lifespan of the technology is realistic?

This question has not been answered yet.

What is the hidden impact of the technology in the whole chain?

This question has not been answered yet.

Now that you have thought hard about the sustainability of this technology, what improvements would you like to make? List them below.

Dit sluit niet aan op onze Ai model.

Future

Did you consider future impact?

This category is only partial filled.

What could possibly happen with this technology in the future?
In de toekomst zou het AI-filter op meerdere sociale media platforms geïmplementeerd kunnen worden om een veiligere online omgeving te creëren. Daarnaast zal de AI steeds slimmer worden, waardoor er minder menselijke input nodig is, behalve voor regelmatige updates

Sketch a or some future scenario (s) (20-50 years up front) regarding the technology with the help of storytelling. Start with at least one utopian scenario.
This question has not been answered yet.

Sketch a or some future scenario (s) (20-50 years up front) regarding the technology with the help of storytelling. Start with at least one dystopian scenario.
This question has not been answered yet.

Would you like to live in one of this scenario's? Why? Why not?
This question has not been answered yet.

What happens if the technology (which you have thought of as ethically well-considered) is bought or taken over by another party?
This question has not been answered yet.

Impact Improvement: Now that you have thought hard about the future impact of the technology, what improvements would you like to make? List them below.
Een verbetering die we zouden toepassen, is dat de AI-filter efficiënter werkt, zodat er minder updates nodig zijn. Dit zou de duurzaamheid en betrouwbaarheid van de technologie vergroten. Daarnaast zouden we een firewall kunnen toevoegen om bestandscorruptie te voorkomen, waardoor de technologie in de toekomst veiliger blijft.