Masterproef industriële ingenieurswetenschappen 2023-2024

De implementatie van een op AI gebaseerde tekst- en afbeeldingsfilter voor het contextafhankelijk modereren van gebruikersinput voor interactieve lichtinstallaties

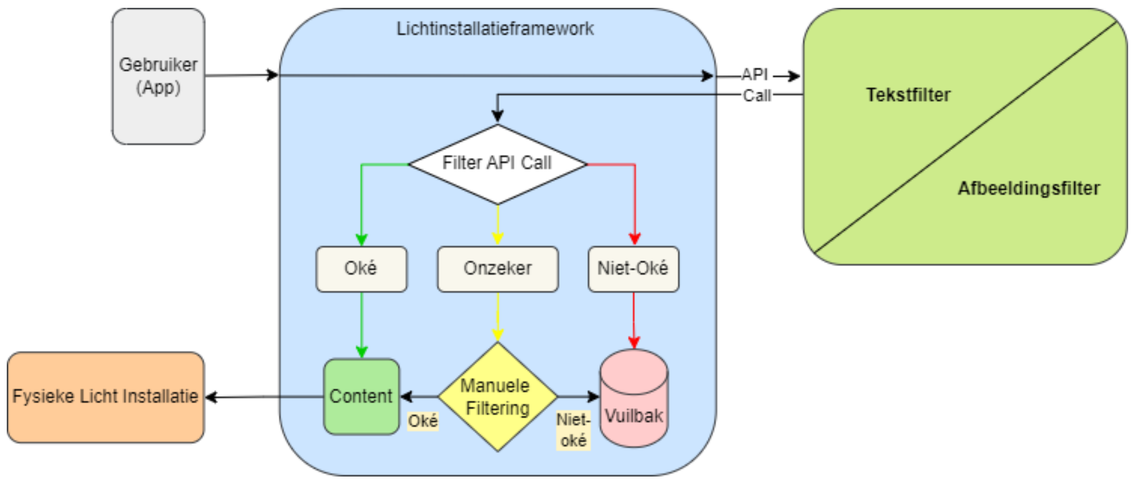
Anneleen Dilissen

Master IW Elektronica-ICT

Probleem

Figuur 1: Voorbeelden interactieve lichtinstallaties [1]

BMK Solutions en Painting with Light maken samen **interactieve lichtinstallaties**. Bezoekers kunnen vrije input sturen naar de interactieve lichtinstallaties in de vorm van **tekst** en **afbeeldingen**, zoals afgebeeld op figuur 1. Niet alle input is echter geschikt om op de installaties te tonen. Daarom is er nood aan een **tekst**- en **afbeeldingsfilter**. Deze filters moeten gebruik maken van AI en aanspreekbaar zijn via een **API**. De filters bepalen wat er met de input moet gebeuren volgens het schema afgebeeld in figuur 2.



Figuur 2: Algemeen schema lichtinstallatieframework

Methode

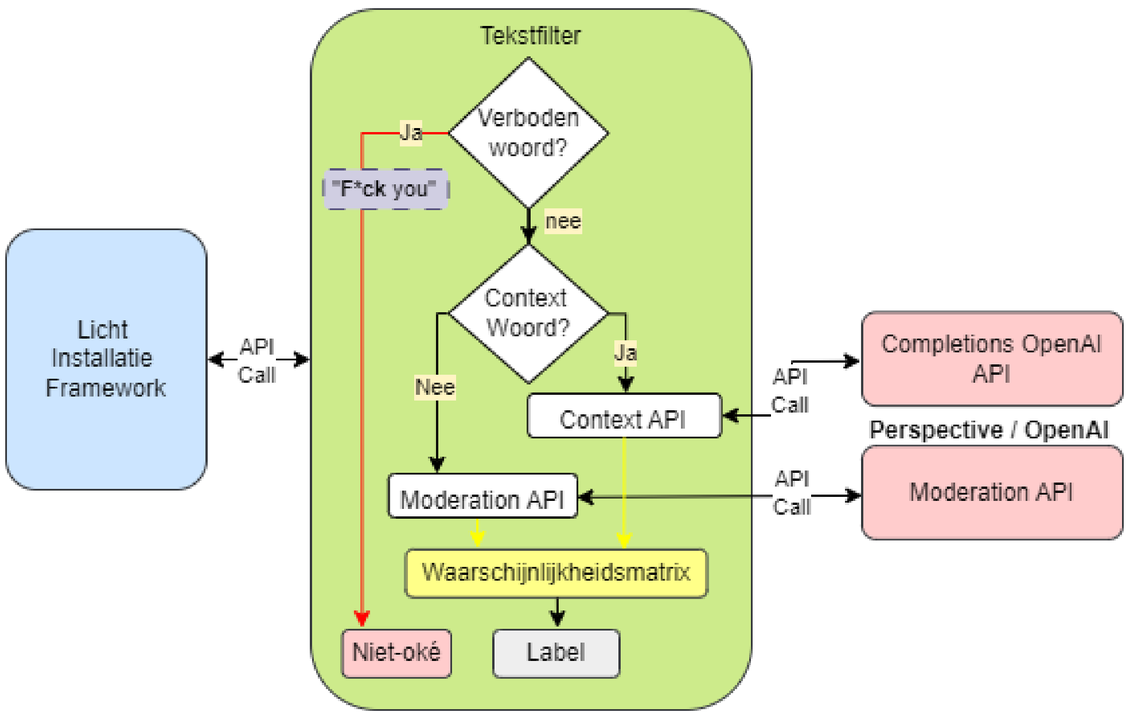
Tekstfilter

De **tekstfilter** bestaat uit vier verschillende modules:

- **Verboden woord** filter en **context** filter
- **Completions module** van OpenAI
- **Perspective** van Google Jigsaw
- **Moderation** module van OpenAI

Deze werken samen zoals in figuur 3.

De eersten **controleren** de tekst op **verboden-** en **context** woorden. De tweede en derde analyseren tekstkenmerken op basis van **waarschijnlijkheidswaarden**. De vierde doet dit ook maar brengt de **context** van de installatie in rekening.



Figuur 3: Schema tekstfilter

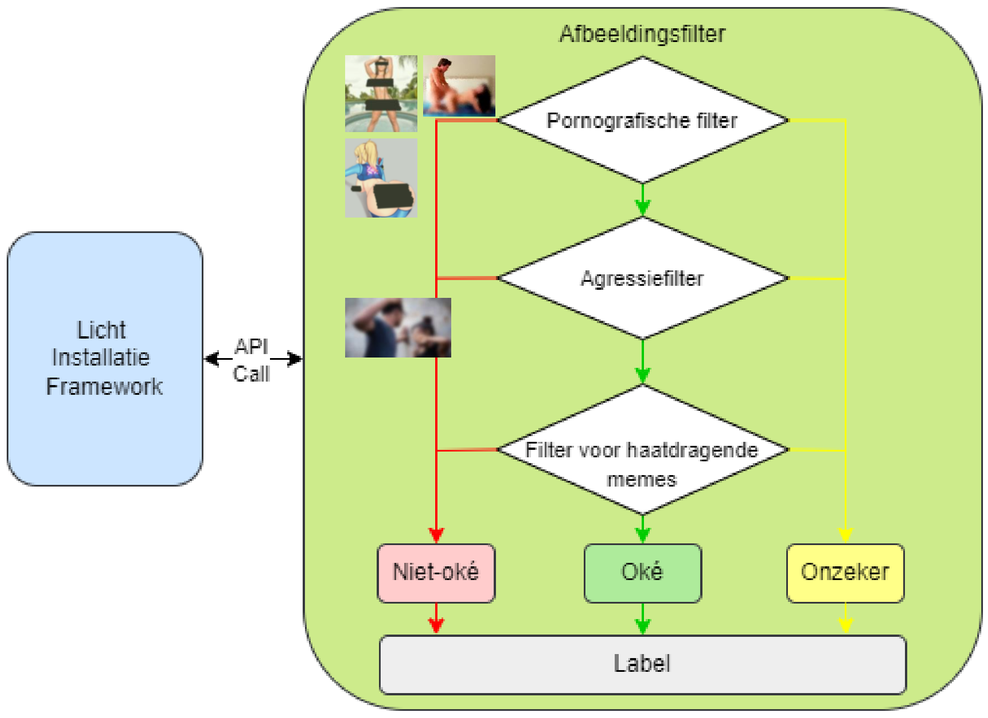
Afbeeldingsfilter

De **afbeeldingsfilter** bestaat uit een *pipeline* van drie modellen:

- de **pornografische** filter
- de **agressiefilter**
- de filter voor **haatdragende *memes***

De *pipeline* wordt afgebeeld in figuur 4.

Het eerste model detecteert **pornografische**, **sexy** en **hentai** afbeeldingen. Het tweede analyseert mogelijke **agressie** in afbeeldingen. Het derde werd wel theoretisch bekeken maar niet geïmplementeerd.



Figuur 4: Schema afbeeldingsfilter

Resultaten

De verboden en context woordfilters zijn getest op **functionaliteit** en blijken **effectief**. De Perspective API en de OpenAI Moderation API werden beide **kort getest** met twee datasets, één van Engelse en één van Nederlandse zinnen. De eerste **presteert goed** bij niet-toegelaten zinnen, terwijl de tweede **inconsistent** is, met name bij Engelse zinnen. De Completions API van OpenAI, gebruikt voor contextuele filtering, toonde **potentieel** maar vereist verdere tests om de effectiviteit te bevestigen.

De afbeeldingsfilter werd **minimaal getest** met een beperkt aantal afbeeldingen per categorie. De **primaire focus** lag op het **functioneel werkend** krijgen van de filter. De opbouw met een ***pipeline*** van verschillende modules blijkt een **veelbelovende aanpak**, omdat het de belasting op latere modellen vermindert. Daarnaast is het hierdoor **eenvoudig** om nieuwe modellen te **integreren** zonder dat hiervoor bestaande modellen aangepast moeten worden.

Conclusie

Een tekstfilter en een afbeeldingsfilter werden geïmplementeerd om de filtering van ongewenste inhoud in tekst en afbeeldingen voor interactieve lichtinstallaties gedeeltelijk te automatiseren. De **accuraatheid** van beide filters hangt af van de **drempelwaarden** ingesteld door de beheerder van de installatie. **Verbetermogelijkheden** en **verder onderzoek** omvatten het verder testen van de **contextuele filtering** met de Completions API, verdere training van het **agressiemodel** met een grote dataset, en de mogelijke **implementatie** van een filter voor **haatdragende memes**.

Promotoren / Copromotoren / Begeleiders

Prof. dr. Kris Aerts
Kevin Moons
Ing. Céline Cuypers

[1] Painting with Light, *Interactieve lichtinstallatiebrochure PWL.* Genk: Painting with Light, 2023.

BMK Solutions PWL PAINTING WITH LIGHT

De opleiding industrieel ingenieur is een gezamenlijke opleiding van UHasselt en KU Leuven

UHASSELT KU LEUVEN