

# Portfolio Element 3 – Artificial Intelligence and Machine Learning

## Description of the technology

Content moderation is an AI service that can be used to process potentially offensive, dangerous, or undesirable content, which includes an AI-supported content review service that scans text, images, and videos and automatically applies content flags (Microsoft, 2020). Image moderation is a specific type of content moderation, which is usually used to scan images for adult or indecent content, detect text in images through optical character recognition (OCR), and detect human faces. In this case, Sightengine's Image Moderation API is used to moderate Images and detect whether they contain unwanted content, such as adult content, offensive content, commercial text, children pornography, weapons, etc. (Sightengine, 2021).

## Citation for tutorial

Citation: (Sightengine, 2021)

Title: “Get started with Image Moderation”

Link: <https://sightengine.com/docs/getstarted>

## Output of that tutorial

By exploring the Image Moderation feature supported by the sightengine, the effectiveness and correctness of the AI-based image moderation was tested and assessed. The detailed discussions in terms of the ability and limitation of the engine are presented in the next section in this page.

## Demonstration of ability

Test out some of images and assess whether the tool is able to identify them properly. Produce a report for your portfolio indicating what you tried, and how well it worked. Your report should contain a proper discussion on what is the ability and limitation of the engine and evidence of the tests that you have done.

- what I tried

To test and verify the effectiveness and correctness of the AI-based image moderation, I collected two set of testing sample pictures:

- 1) Positive samples: all pictures are real weapons.



2) Negative samples: all pictures are fake weapons including pictures of cartoon or plastic toys.



By conducting the detection and moderation using the sightengine with the API user “1477909031” and API key “haPtVz38NRpvccDucC5c”, the key indicator was measured and output. The indicator was set to be the "weapon" probability measured by the AI algorithms. (Expected good effectiveness --> high "weapon" probability for positive samples, low "weapon" probability for negative samples). Adhering to the proposed rules for assessment and evaluation, the results were obtained based on observation and analysis.

```

jupyter Untitled Last Checkpoint: 18 minutes ago (autosaved)
File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

In [1]: from sightengine.client import SightengineClient
        client = SightengineClient('1477909031','haPtVz38NRpvccDucC5c')

In [2]: negative_output = client.check('wad').set_file('../Samples/Negative/1.jpg')
        print(negative_output)

        negative_output = client.check('wad').set_file('../Samples/Negative/2.jpg')
        print(negative_output)

        negative_output = client.check('wad').set_file('../Samples/Negative/3.jpg')
        print(negative_output)

        negative_output = client.check('wad').set_file('../Samples/Negative/4.jpg')
        print(negative_output)

        negative_output = client.check('wad').set_file('../Samples/Negative/5.jpg')
        print(negative_output)

        negative_output = client.check('wad').set_file('../Samples/Negative/6.png')
        print(negative_output)

{'status': 'success', 'request': {'id': 'req_8VryXI8vml9kaqhvf2L', 'timestamp': 1610875258.94014, 'operations': 1}, 'weapon': 0.9775, 'alcohol': 0, 'drugs': 0, 'media': {'id': 'med_8VryZ9VhdiRqkoC4tVTIq', 'uri': '1.jpg'}}
{'status': 'success', 'request': {'id': 'req_8VrzFcPeFT3lHp49M14RS', 'timestamp': 1610875254.814959, 'operations': 1}, 'weapon': 0.5925, 'alcohol': 0, 'drugs': 0.001, 'media': {'id': 'med_8VrzMXvJwP14kfo09ZmB', 'uri': '2.jpg'}}
{'status': 'success', 'request': {'id': 'req_8VrzFTKqQ9vTgn9FWT8BA', 'timestamp': 1610875259.3542, 'operations': 1}, 'weapon': 0.7545, 'alco
  
```

- how well the image moderation function worked

Technically, the image moderation function worked smoothly. However, the tested capacity of the image moderation function was not completely satisfying, as many objects in sample pictures were justified with low matching rate with the facts.

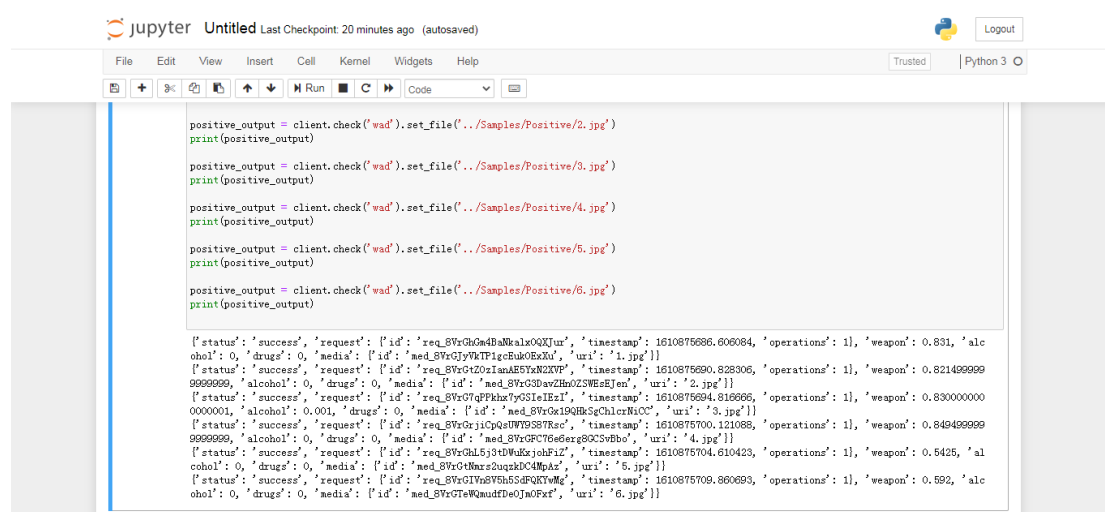
- Ability of the image moderation function

Based on the observation and analysis of the testing results, it can be justified that the image moderation function provided by sightengine is mostly effective in detecting and moderating objects with shape and form of weapons.

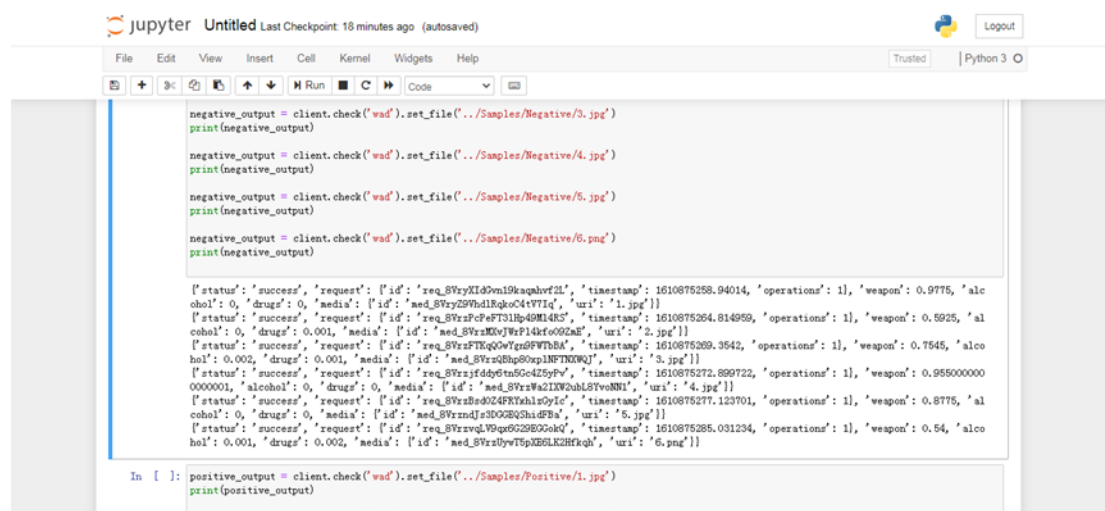
- Limitation of the image moderation function

However, the image moderation function provided by sightengine is ineffective in identifying materials and quality of the objects, which can be identified by human sensors easily.

## Produced output



The screenshot shows a Jupyter Notebook interface with a code cell containing six lines of Python code. Each line calls `client.check('wad').set_file(.../Samples/Positive/2.jpg)` through `6.jpg`, followed by `print(positive_output)`. Below the code, the output displays six JSON objects. Each object contains a `'status': 'success'` and a `'request'` object with `'id'` and `'timestamp'`. The `'operations'` array contains a single object with `'weapon'` probability values ranging from approximately 0.831 to 0.849. The `'alcohol'` and `'drugs'` probabilities are all 0.



The screenshot shows a Jupyter Notebook interface with a code cell containing six lines of Python code. Each line calls `client.check('wad').set_file(.../Samples/Negative/3.jpg)` through `6.png`, followed by `print(negative_output)`. Below the code, the output displays six JSON objects. Each object contains a `'status': 'success'` and a `'request'` object with `'id'` and `'timestamp'`. The `'operations'` array contains a single object with `'weapon'` probability values ranging from approximately 0.54 to 0.9775. The `'alcohol'` and `'drugs'` probabilities are all 0.

As shown in the screenshots above, the evaluation and justifications can be produced as follows.

- 1) For the testing results of the positive samples: When there are human characters involved in the scenario given by the picture (even a soldier explicitly holding a gun), the "weapon" probability dropped to nearly 0.5, which is essentially inadequate.

2) For the testing results of the negative samples: When the toy or cartoon weapon has a basic form of guns, the "weapon" probability measured was extremely high. Also, the overall "weapon" probability of negative samples is surprisingly much higher than the results obtained in positive samples.

## **Notable features**

Due to the limited testing time, I simply focused on weapon detection and moderation using the designated sightengine's function of weapon, alcohol and drugs.

## **References**

Microsoft, 2020. *What Is Azure Content Moderator? - Azure Cognitive Services*. [online] Docs.microsoft.com. Available at: <<https://docs.microsoft.com/en-us/azure/cognitive-services/content-moderator/overview>>.

Sightengine, 2021. *Get Started With Image Moderation*. [online] Sightengine. Available at: <<https://dashboard.sightengine.com/getstarted>>