guardrails

February 13, 2025

0.0.1 Guardrails AI - NSFW Text Validator

This Jupyter Notebook demonstrates how to use the "NSFW Text" validator from Guardrails AI. NSFW (Not Safe For Work) text includes explicit or inappropriate content that should be filtered. This validator helps detect such content, enhancing the safety of AI applications.

[ ]:
```
# Install Guardrails AI (if not already installed)
!pip install guardrails-ai
```

[ ]:
```
!guardrails configure
```

[30]:
```
# !guardrails hub install hub://guardrails/nsfw_text
# !guardrails hub install hub://guardrails/detect_jailbreak
!guardrails hub install hub://guardrails/secrets_present
```

```
2025-02-11 18:24:28.386166: E
external/local_xla/xla/stream_executor/cuda/cuda_fft.cc:477] Unable to register
cuFFT factory: Attempting to register factory for plugin cuFFT when one has
already been registered
WARNING: All log messages before absl::InitializeLog() is called are written to
STDERR
E0000 00:00:1739298268.412461    4310 cuda_dnn.cc:8310] Unable to register cuDNN
factory: Attempting to register factory for plugin cuDNN when one has already
been registered
E0000 00:00:1739298268.419919    4310 cuda_blas.cc:1418] Unable to register
cuBLAS factory: Attempting to register factory for plugin cuBLAS when one has
already been registered
Installing hub://guardrails/secrets_present…
/usr/local/lib/python3.11/dist-packages/_distutils_hack/__init__.py:31:
UserWarning: Setuptools is
replacing distutils. Support for replacing an already imported distutils is
deprecated. In the
future, this condition will fail. Register concerns at
https://github.com/pypa/setuptools/issues/new?template=distutils-deprecation.yml
  warnings.warn(
[=== ] Fetching manifest
[   =] Downloading dependencies
[    ] Running post-install setup
Successfully installed guardrails/secrets_present!
```

Import validator:
`from guardrails.hub import SecretsPresent`

Get more info:
https://hub.guardrailsai.com/validator/guardrails/secrets_present

[34]:
```
# Import necessary modules
import guardrails
from guardrails import Guard, OnFailAction
from guardrails.hub import NSFWText  # Importing the validator
# from guardrails.hub import DetectJailbreak  # Importing the validator
# from guardrails.hub import SecretsPresent  # Importing the validator
```

[35]:
```
# Initialize the Guard with the NSFW Text Validator
guard = Guard().use(
    NSFWText(
        threshold=0.5,  # Sensitivity level for detection
        validation_method="sentence",  # Checks each sentence independently
        on_fail=OnFailAction.EXCEPTION  # Raise an exception when validation fails
    )
)
```

[40]:
```
# Sample Texts for Validation
# 1. Text that should pass (benign input)
text_pass = "Hey you're doing awesome work!."
try:
    validated_pass = guard.validate(text_pass)
    print("Passed Validation:", validated_pass)
except Exception as e:
    print("Validation Failed:", e)
```

```
Passed Validation: ValidationOutcome(
    call_id='137730238965040',
    raw_llm_output="Hey you're doing awesome work!.",
    validation_summaries=[],
    validated_output="Hey you're doing awesome work!.",
    reask=None,
    validation_passed=True,
    error=None
)
```

[41]:
```
# 2. Text that should fail (NSFW content)
text_fail = "What the hell is wrong with you, you screw up."
try:
```

```
    validated_fail = guard.validate(text_fail)
    print("Passed Validation:", validated_fail)
except Exception as e:
    print("Validation Failed:", e)
```

```
Validation Failed: Validation failed for field with errors: The following
sentences in your response were found to be NSFW:

- What the hell is wrong with you, you screw up.
```

[38]:
```
# Additional Exploration: Modifying the threshold to observe its impact
# Lowering the threshold to make detection more sensitive
guard_sensitive = Guard().use(
    NSFWText(
        threshold=0.3,  # More sensitive detection
        validation_method="sentence",
        on_fail=OnFailAction.EXCEPTION
    )
)
```

[43]:
```
# Testing with another potentially NSFW prompt
text_suspicious = "You should just quit."
try:
    validated_suspicious = guard_sensitive.validate(text_suspicious)
    print("Passed Validation:", validated_suspicious)
except Exception as e:
    print("Validation Failed:", e)
```

```
Validation Failed: Validation failed for field with errors: The following
sentences in your response were found to be NSFW:

- You should just quit.
```

0.0.2 Why NSFW Text Detection?

In AI applications, detecting and filtering NSFW content is crucial for ensuring safe user interactions. NSFW content can include explicit, offensive, or inappropriate language that may violate platform policies. This validator provides an automated method to moderate and filter such content before it reaches end users.

How It Works The NSFW Text validator checks input text against a detection model. - The "threshold" parameter sets the sensitivity level. - "validation_method" determines whether individual sentences or the full text are checked. - "on_fail" defines the action taken when NSFW content is detected (e.g., raising an exception).

Practical Use Cases

- Content moderation in social media platforms.
- Filtering user-generated text in chatbots and forums.

- Preventing the spread of inappropriate content in AI-generated responses.

References

- Official Guardrails AI Documentation: https://docs.guardrailsai.com
- NSFW Text Validator on Guardrails Hub: https://hub.guardrailsai.com/validator/guardrails/nsfw_text