spectrum labs

Custom Detection List

User Guide

spectrum labs

Custom Detection List

What is it?

Spectrum is adding a new solution to every customer's configuration in order to give you a quick way to add coverage for terms that are not currently being flagged.

Edits and additions to this list may happen outside of scheduled releases and as often as weekly.

Results from this custom detection list will be returned as a separate solution within the API Response.

Sample Output

The API response will include a boolean determination for a solution called "custom". Here is what that looks like:

```
{
  "contentId": "some-id-for-the-message",
  "behaviors": {
    "profanity": true,
    "insult": false,
    "sexual": false,
    "csam": false,
    "radicalization": false,
    "custom": true
  }
}
```

spectrum labs

Use Cases

This list can be useful for a variety of scenarios. These include:

- Anything that may be missed by the current solutions but is important for a customer to be able to flag
- A Zero Tolerance list for terms that may or may not be covered by other solutions' definitions
- Terms that are outside of the scope of the licensed behavior's definitions
- Known bad terms in languages outside of what was licensed

How this works

When you have a new term that you would like to start flagging as part of this solution, start by telling your Customer Success team. They will review the request and add it to the "custom" solution as part of the next push.

The "custom" solution will be updated every Tuesday when there are new additions or edits.

For Example

Let's say you wake up on a Monday morning and start seeing references to "#letsgobrandon" on your platform. You realize this is a masked profane statement coming out of a recent world event and you want to make sure it gets flagged. If you tell your Spectrum team right away, they will work to include it in Tuesday's push.

A few things to note

Because this solution falls outside of your licensed Spectrum behaviors the same model performance SLAs do NOT apply.

This solution is designed to return a true whenever it sees an exact match for the terms that are included. There is no accounting for context, detected language, or special pre-processing for leet speak at this time. This is also focused on exact matches so these will all be token matches, not sub-strings. As a result if "moon" is added to this list "the man in the moon" would match but "Moonchild is the name given to the childlike empress" would not.