spectrum labs

Confidence Buckets

User Guide

spectrum labs

Confidence Buckets

What are they?

Spectrum is adding a new set of fields to the API Response in order to give customers additional signal to use when making decisions about how to take action.

These fields are correlated to different confidence buckets which reflect how confident the model is that the behavior is present in the data.

At initial roll out there will be three different values:

Value	Description
High	This value is returned when the models are the most confident that the data matches the behavior it is identifying. Results from this bucket should have the highest precision/lowest false positive rate.
Low	This value is returned when the model thinks the data matches the behavior it is identifying but is less confident. This is still above the threshold for a "True" determination but is expected to have a slightly lower precision/higher false positive rate.
NotDetected	This value is returned when the model does not think the data matches the behavior. This is when the criteria for the solution is not met - this includes scores that are below the threshold for determination.

*Note: Medium may be added as an additional bucket in future releases.

These will be returned per behavior in addition to the per behavior Boolean determinations.

spectrum labs

Sample Output

The new API response will look like this:

```
{
  "contentId": "some-id-for-the-message",
  "behaviors": {
    "hate-speech": true,
    "insult": false,
    "sexual": false,
    "profanity": true,
    "bullying": true,
    "custom": true
  },
  "confidences": {
    "hate-speech": "High",
    "insult": "NotDetected",
    "sexual": "NotDetected",
    "profanity": "Low,
    "bullying": "Low",
    "custom": "Low"
  },
  "language": "eng"
}
```

spectrum labs

How to use

The introduction of this new field gives you additional signal that can be used to update your Behavior Action Matrix.

For example, here is an idea for how you might want to incorporate confidence for Hate Speech results:

Behavior Identification	Confidence	Action
True	High	Automate Redaction
True	Low	Create a Case for Review
False	NotDetected	N/A

Note: Because this is a new field you will need to update your filtering capabilities in the technical integration to expand your real time redaction capabilities.

A few things to note

Confidence does NOT equal severity

High confidence means that the model is very sure that the behavior is present in the data. It does not mean that the behavior is more severe than say a Low confidence.

While all behaviors will have confidence returned, some will only return the default response

Spectrum will turn this feature on with the default value of "Low" for all behaviors. Each behavior will then be configured to have a delineation between "High" and "Low" over the next couple of releases. Clear communication around which behaviors are using the default "Low" and which have been configured to have a delineation will be included in each production push's release notes.

The first behavior to be fully configured for all buckets will be Hate Speech.