besedo

8 steps to accurate content moderation filter creation

Looking for a headstart?

1. Mission

- [ ] Define the goal of the filter.
- [ ] Create the filter in Implio and give it a descriptive name.

2. Local adjustment

- [ ] Check laws and regulations of the country you're operating in.
- [ ] Adjust the filter to adhere to cultural and national laws and sensitivities in the region.

3. Action

- [ ] Decide on the action you want the filter to take (refuse, send to manual or no action).
- [ ] At this step you can also set up test filters with no action other than highlighting ads that would have been caught.

4. List

- [ ] Create a list of all keywords related to what you want to catch (e.i. for drugs, add cocaine, heroin, cannabis etc. to the list).
- [ ] Add any slang words that are likely to be used for what you want to catch.

5. Rule

- [ ] Set up your rule, and make sure that it utilizes any relevant lists you've created.
- [ ] Add exceptions to avoid false positives.

6. Quality control

- [ ] Perform the first quality check.
- [ ] Compile a list of any false positives for further investigation.

7. Exceptions

- [ ] Add exceptions for all false positives (so called white-listing of specific content).

8. Rinse and repeat

- [ ] Run your data through again (including exceptions) to quality check your updated filter.
- [ ] Repeat step 6-8 as many times as you have to in order to reach your target quality rate. At Besedo we aim for 95% accuracy as a minimum and most of our filters reach higher levels.

besedo

Looking for a headstart?

Here are some resources:

Counterfeit filter checklist

How to create accurate content moderation filters that work

About us

At besedo we have been working with user generated content moderation since 2002 with the aim to help sites provide better digital experiences.

We offer data-driven moderation services and solutions through AI moderation, automated filters and manual moderation.

hi@besedo.com

besedo.com

The besedo blog

@besedo_official

@besedo

@Besedo