twitch.json

hatespeech

Page ID: 20

Source: https://safety.twitch.tv/s/article/Chat-Tools?language=en\_US

['discrimin', 'hate']

AutoMod AutoMod uses machine learning and natural language processing algorithms to hold potentially inappropriate or offensive messages from chat so they can be reviewed by the creator or a channel moderator before appearing to other viewers in the chat.

With AutoMod, creators can establish their own reliable baseline for acceptable language and around-the-clock chat moderation, a way to create a more positive experience for themselves and their communities.

When a user sends a message that AutoMod flags as potentially inappropriate, the message is held pending approval or denial.

If approved, it will be released to chat.

Creators can configure AutoMod by selecting one of four levels.

These levels affect how aggressively AutoMod holds back messages it suspects are risky within four categories: discrimination, sexual content, hostility, and profanity.

These settings can only be changed by the broadcaster, while the dictionary each setting contains is managed by Twitch.

You can find more information about AutoMod in our "How to Use AutoMod" help article .

Blocked/Permitted Terms You can also limit the use of certain words or phrases in your chat, or allow certain words to bypass AutoMod's filters.

This works when AutoMod is turned on and when it is turned off.

Blocked Terms When adding a word or phrase to your blocked items, you can choose to set it to “Public” or “Private.”

Setting a term to Private; will prevent it from being viewable to your channel's moderators.

This is recommended for terms that involve personally identifiable information that you'd rather keep private.

Once you have added a term, you can click the pencil icon to edit it, or the trash icon to remove it from your blocked terms list.

When adding blocked terms, you can also use wildcards by adding a "\*" onto your term.

For example, adding "hate\*"; to your blocked terms will also cause the term "haters" and "hateful" to be blocked.

You can also use this feature to filter any variant of a url.

Permitted Terms Permitted terms are used to set exceptions for specific words or phrases and bypass AutoMod's filters.

Adding terms or phrases to your permitted terms will prevent AutoMod from catching messages containing those terms (provided those messages do not contain other prohibited language, either flagged by AutoMod or your blocked terms).

For example, if AutoMod catches an innocuous phrase as a false positive, you can manually add this phrase to your permitted terms so that it is not flagged again.

Alternatively, clicking “Allow” on a message held by AutoMod will temporarily add the caught phrase or word to your permitted terms.

Repeatedly allowing a word or phrase will increase the amount of time it is permitted from one hour, to a day, to seven days, to permanently.