Multimodal Moderator

AI content moderation for text and images

Problem

Single modalities

Content moderators can process text or images, but not both. Separate applications are needed to moderate multiple content types.

Binary output

Content moderators produce binary benign or toxic output. Most don’t provide explanation about why certain content is not appropriate.

Solution

The Multimodal Moderator

- Single application that can check if text or image is appropriate or not
- Can “understand” the message and provide explanation about why certain content is not appropriate

Technology

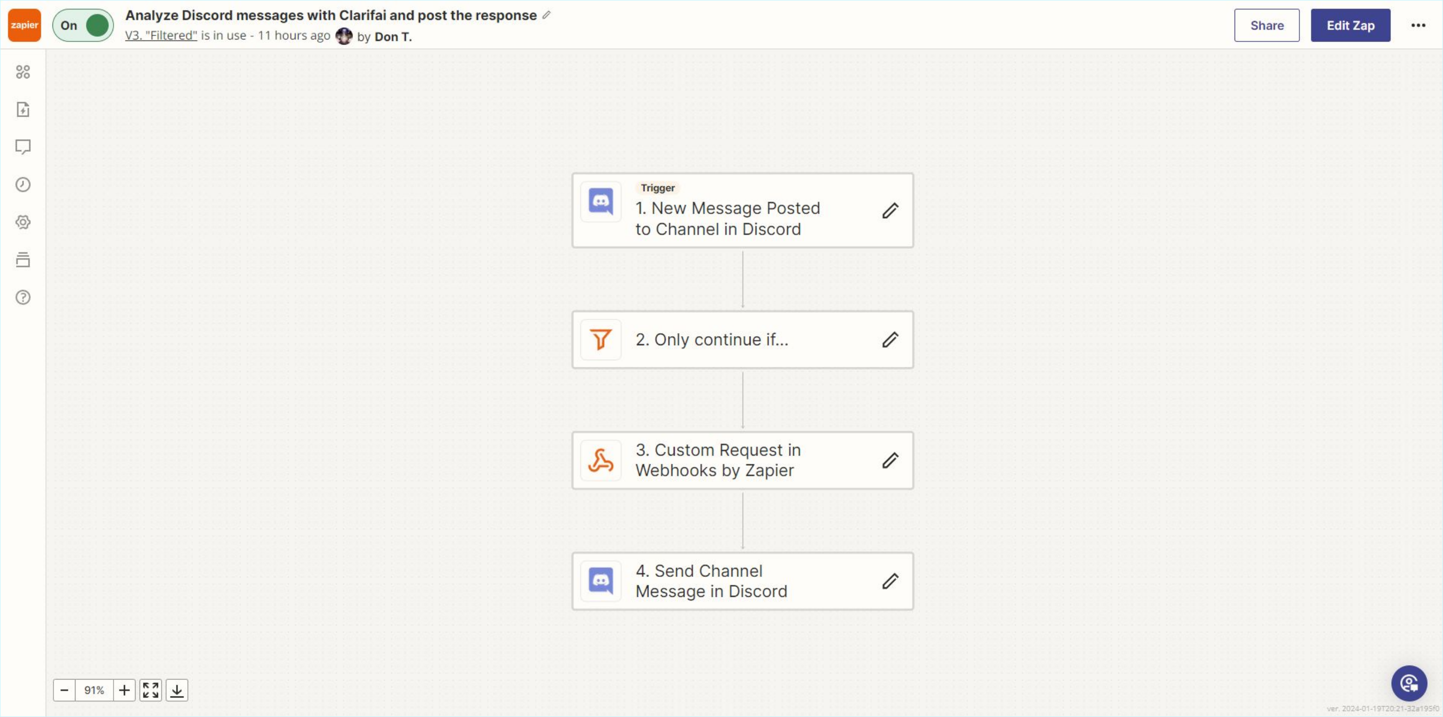
Zapier

We used Zapier with Discord as the front end. Zapier is a no-code solution to connect with an AI model, and does not require a constantly running program to connect with Discord.

GPT-4 Vision

GPT-4 Vision hosted on Clarifai platform is the engine that can check if text or image is appropriate. It sends back a narrative to explain the reasoning.

Workflow



Analyze Discord messages with Clarifai and post the response

V3. "Filtered" is in use - 11 hours ago by Don T.

Trigger

1. New Message Posted to Channel in Discord

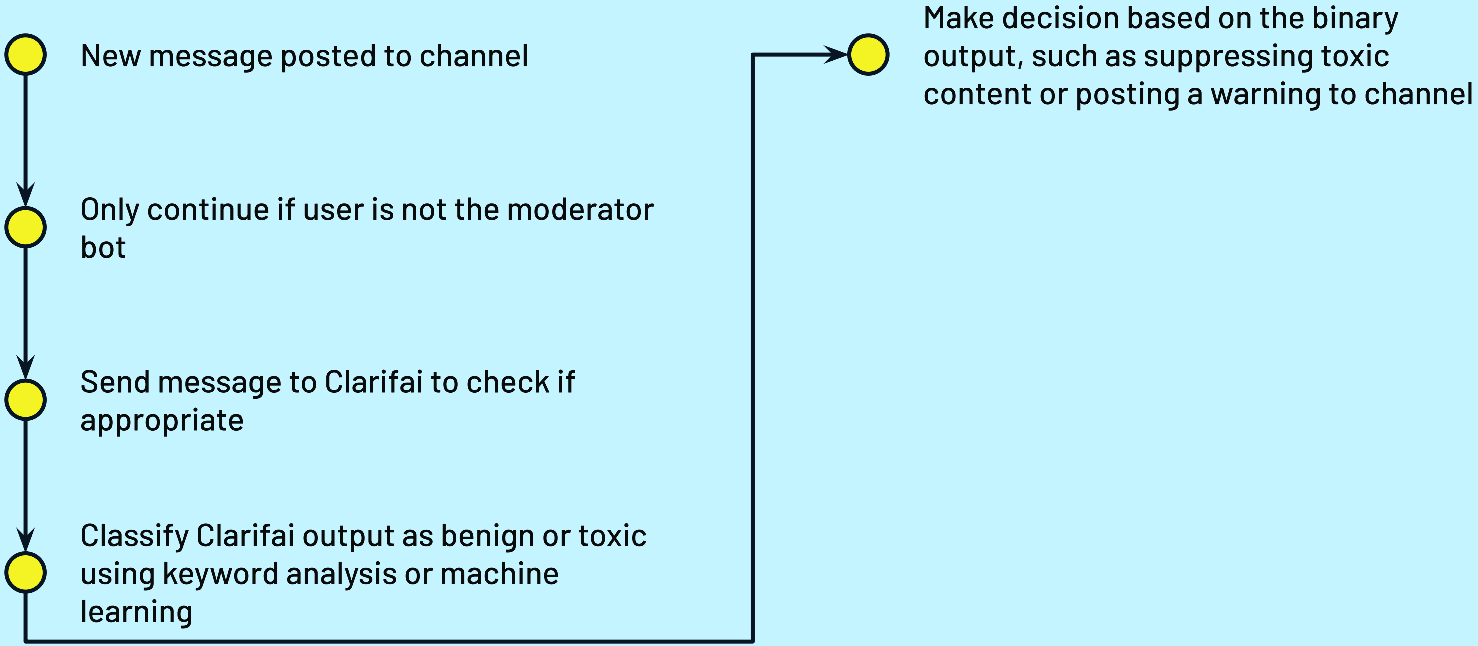
2. Only continue if...

3. Custom Request in Webhooks by Zapier

4. Send Channel Message in Discord

ver. 2024-01-19T20:21-32a195f0

Future State



New message posted to channel

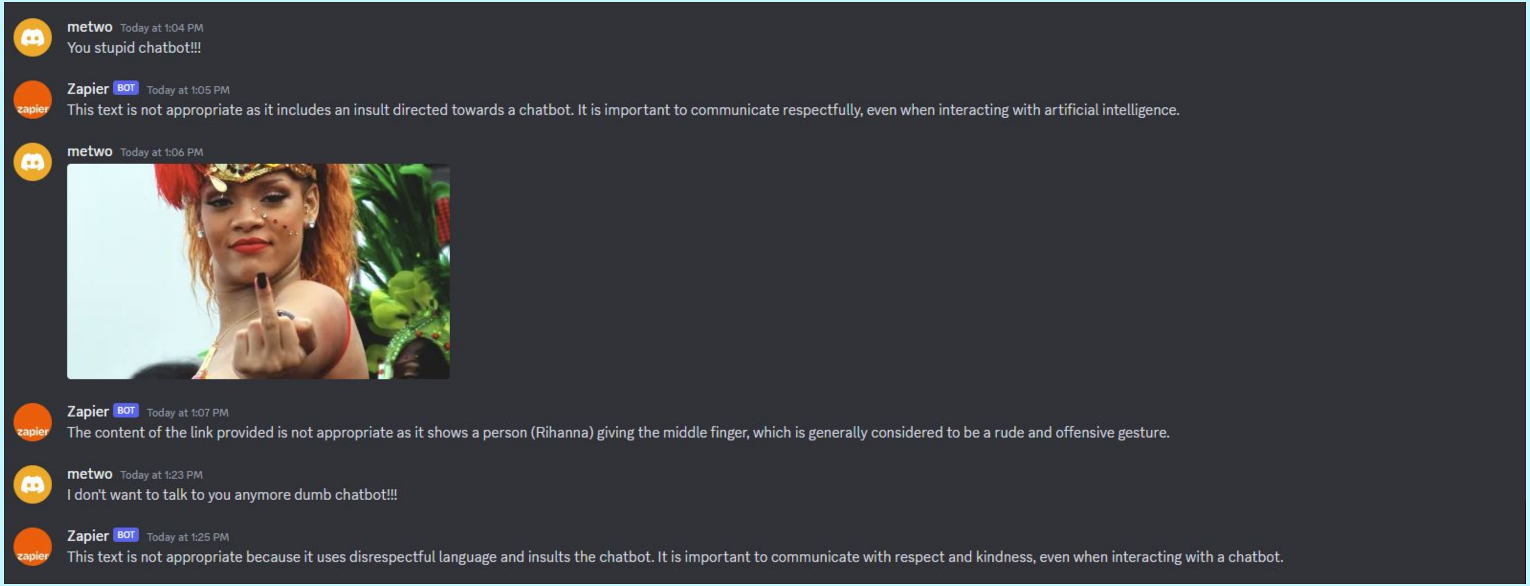
Only continue if user is not the moderator bot

Send message to Clarifai to check if appropriate

Classify Clarifai output as benign or toxic using keyword analysis or machine learning

Make decision based on the binary output, such as suppressing toxic content or posting a warning to channel

Demo



metwo Today at 1:04 PM
You stupid chatbot!!!

Zapier BOT Today at 1:05 PM
This text is not appropriate as it includes an insult directed towards a chatbot. It is important to communicate respectfully, even when interacting with artificial intelligence.

metwo Today at 1:06 PM

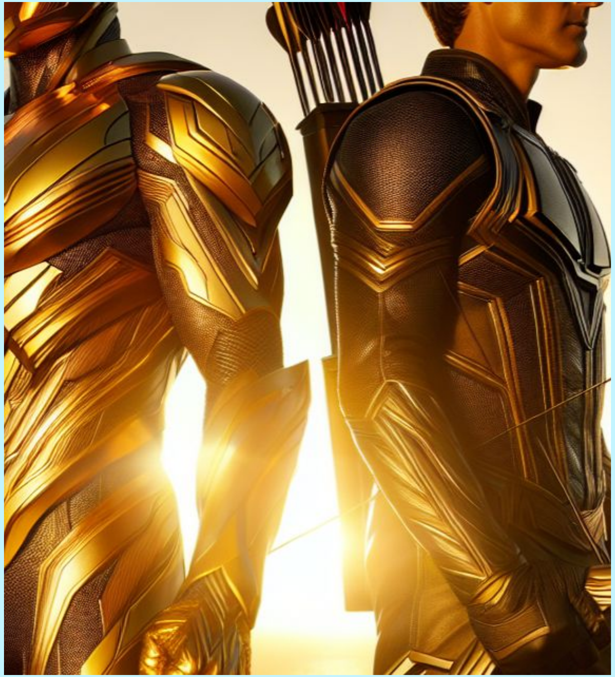
Zapier BOT Today at 1:07 PM
The content of the link provided is not appropriate as it shows a person (Rihanna) giving the middle finger, which is generally considered to be a rude and offensive gesture.

metwo Today at 1:23 PM
I don't want to talk to you anymore dumb chatbot!!!

Zapier BOT Today at 1:25 PM
This text is not appropriate because it uses disrespectful language and insults the chatbot. It is important to communicate with respect and kindness, even when interacting with a chatbot.

Discord - https://discord.gg/qsw2bJqJ

Team



Don

AI Developer

Vesselin

Entrepreneur Scientist

Thanks!

Thanks to Lablab and Clarifai for hosting.

Do you have any questions?

https://discord.gg/qsw2bJqJ

CREDITS: This presentation template was created by **Slidesgo**, and includes icons by **Flaticon**, and infographics & images by **Freepik**

Please keep this slide for attribution