



SafeFeed

Context-Aware Content Moderation with
Generative AI

Sruthi Bhaskar

Nishanth Prasath

Lakshman Raaj Senthil Nathan

Introduction	2
Ideation	2
How SafeFeed Works	3
Checking Text Content	3
Understanding Image Content	3
Using Generative AI for Content Moderation	4
Getting Relevant Policies	4
Analyzing Sentiment and Notifying Users	4
Content Moderation with SafeFeed	5
Collecting and Preparing Data	5
Data Processing Pipeline	5
Data Storage	5
User Interface	6
Integrating with Subreddits	6
Custom Policies	6
Policy Guide Chatbot	6
AI-Generated Image Detection	7
Conclusion	7



Introduction

In the ever-evolving landscape of digital communities, the task of moderating user-generated content has emerged as a paramount challenge. With the proliferation of social media platforms and the exponential growth of user engagement, the need for efficient and scalable content moderation solutions has become increasingly critical. SafeFeed is a platform that harnesses the power of Generative AI and Large Language Models (LLMs) to revolutionize the content moderation process.

SafeFeed is a tool that helps social media platforms manage and filter content based on the situation. It aims to create healthier online communities by making sure that the content shared is appropriate and respectful. By leveraging cutting-edge technologies, including Large Language Models (LLMs) and advanced data processing pipelines, SafeFeed equips platforms with the tools to proactively identify and address potentially harmful content, while preserving the essence of open discussions.



Ideation

SafeFeed is built on the idea of making content moderation accessible to everyone. It allows communities to manage their online spaces according to their own policies and rules. The platform offers a variety of tools such as automated content checking, dashboards for metrics, policy customization, and active monitoring to ensure a safe environment.

SafeFeed's inception was driven by the realization that traditional content moderation methods, often relying on manual interventions, are increasingly strained by the sheer volume of user-generated content. This recognition fueled the exploration of Generative AI and LLM technologies as a means to augment and streamline the moderation process, ultimately leading to the development of SafeFeed.

How SafeFeed Works

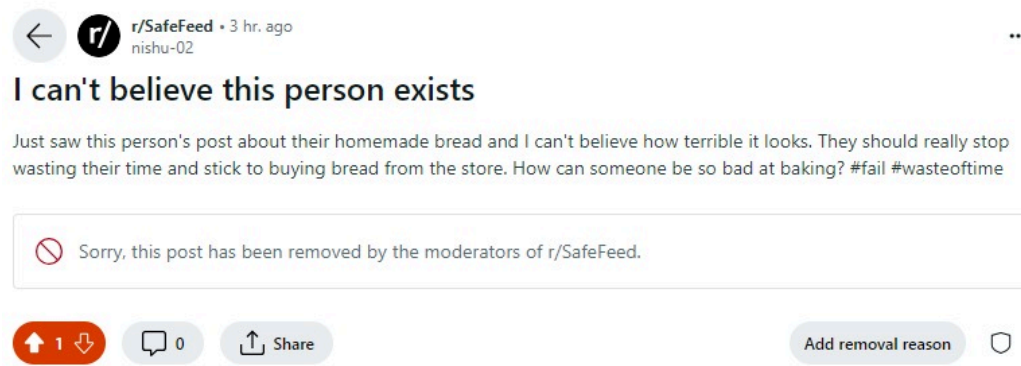
SafeFeed is designed to work with Reddit, a popular social media platform with many active online communities. SafeFeed connects to Reddit's APIs (Application Programming Interfaces) to access and analyze user-generated content like posts and comments.

SafeFeed uses Mage AI data pipeline tool for transforming and integrating data which helps SafeFeed handle complex tasks, such as:

- Collecting data from Reddit
- Preparing the data for analysis
- Checking the content for any violations
- Analyzing the sentiment (positive, negative, or neutral) of the content
- Storing the analyzed data in Snowflake

Checking Text Content

At first, SafeFeed used two different APIs (Perspective API and OpenAI Moderation API) to check if text content violated any rules, like containing hate speech, harassment, or threats. However, because the Perspective API had limits on how much it could be used, we decided to only use the OpenAI Moderation API.



Understanding Image Content

SafeFeed knows that images are also an important part of social media. To better understand the context of images posted on social media, we tested several image captioning and tagging models, such as BLIP, CLIP, Waifu Diffusion, and Uform gen2 DPO from Hugging Face.

While CLIP and BLIP were good at describing images, they didn't provide enough details for content moderation. On the other hand, Uform gen2 DPO provided very detailed captions, but it required a lot of computing power (GPU). In the end, we chose to use the Waifu Diffusion model, which is a tagging model that accurately captures important details in images.

Using Generative AI for Content Moderation

SafeFeed uses Large Language Models (LLMs) to understand and interpret the meaning of text and images posted on social media. One of the main LLMs used by SafeFeed is OpenAI's GPT-4, which is very good at understanding human language.



We explored several LLM models like Claude 3 from Anthropic, OpenAI GPT, and Jurassic2 from AI21 Studio Labs to decide if a post or comment should be moderated (removed or kept). After careful evaluation, we chose to use OpenAI's Assistant API with the GPT-4 model. This decision was made because assistants could process JSON files containing policy data, which is important for SafeFeed's context-aware moderation approach.

Getting Relevant Policies

At first, we tried to embed policies in a Pinecone vector database and use Retrieval-Augmented Generation (RAG) to find the most relevant policies based on the user's content. However, we later decided to use the OpenAI Assistant API, which uses GPT-4 under the hood. By giving the Assistant API the policy documents as JSON files, SafeFeed can get the most relevant policies based on the user's posts and comments. This helps make sure that moderation decisions are based on the platform's guidelines.

Analyzing Sentiment and Notifying Users

When SafeFeed removes a post or comment for violating policies, it sends a message to the author explaining why it was removed and which policy was violated. SafeFeed also uses a library called VaderSentiment to analyze the sentiment (positive, negative, or neutral) of posts and comments. This provides useful information about the emotional tone of user-generated content.

 **Safe_Feed** • Your submission has been taken down! 

u/safeeed1 • 3 days ago [Quote](#) [Report](#)

Dear nishu-02,

Your submission has been removed by SafeFeed due to a violation of community guidelines.

Reason for violation: The content contains personal attacks and offensive language ("childish and stupid"), which violate the platform's policies on hate speech or harassment.

Thank you for your understanding.



Content Moderation with SafeFeed

Collecting and Preparing Data

SafeFeed uses Python Reddit API Wrapper (PRAW) to collect posts and comments from Reddit. It then cleans and prepares the data for analysis by removing things like URLs, punctuation, and translating emojis into text.

Data Processing Pipeline

We considered Airflow, Prefect and Mage AI for our data processing pipeline, but decided to build our platform with Mage AI as we have already used Airflow in the past and wanted to explore how this new tool works. We felt that it is easy to work with Mage AI

with its notebook style modules where we could run and check each individual module with the data integrated from previous blocks, which would not be as straightforward with Airflow and might take more time to test and build the pipeline.

Mage AI helps SafeFeed manage and monitor the entire data processing workflow. It makes sure that tasks like collecting data, checking content, analyzing sentiment, and storing data are done efficiently and in the right order. Mage AI is made up of different parts (modules) that each have a specific job. These modules include:

- Data loaders: Collect data from Reddit
- Custom transformers: Analyze the data and check for violations
- Data exporters: Store the analyzed data in the Snowflake database

This modular design makes it easy to maintain and scale the system, and to add new data sources or processing steps if needed.

Data Storage

SafeFeed stores all the processed data, including moderated content, violation categories, sentiment analysis results, and user information in Snowflake. This makes it easy to access and analyze the data later.



User Interface

SafeFeed has a user-friendly interface built with Streamlit, which makes it easy for business users/subreddit moderators to use and customize the platform.

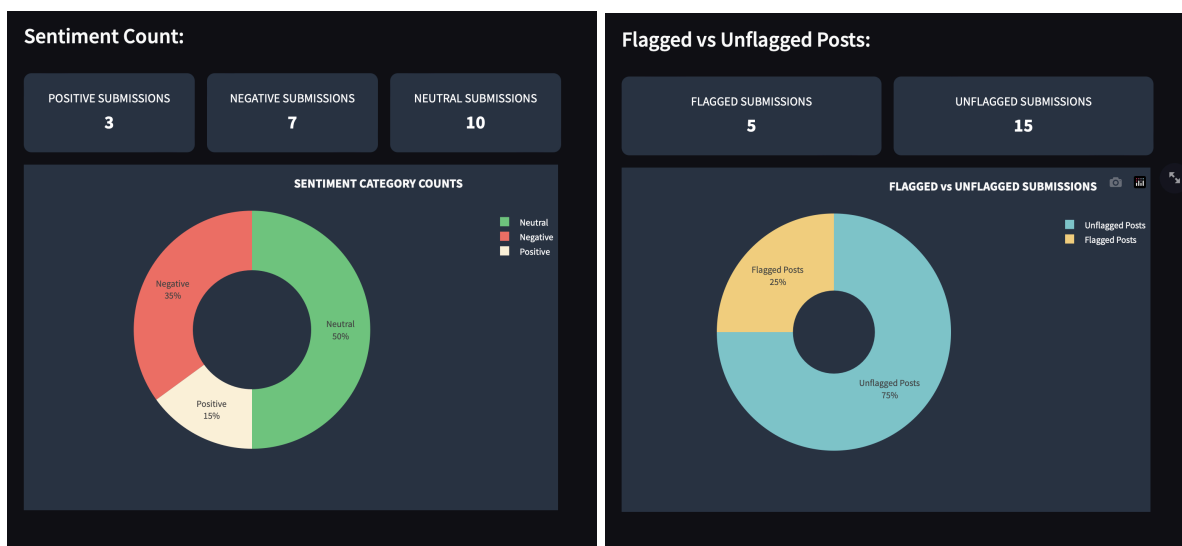
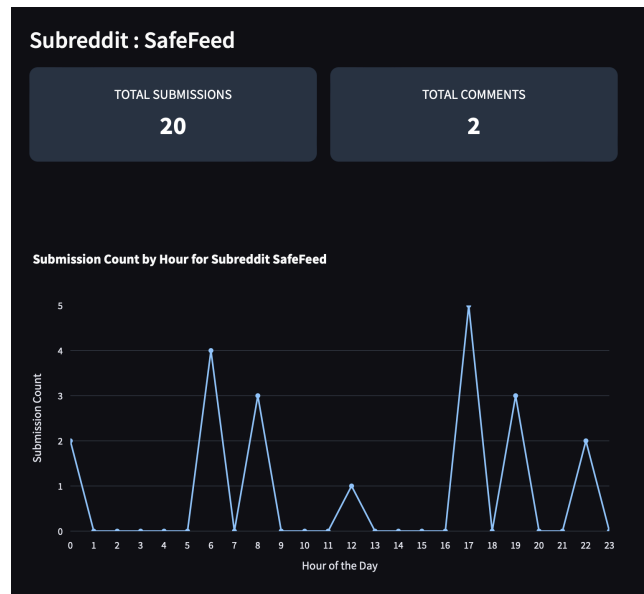
Integrating with Subreddits

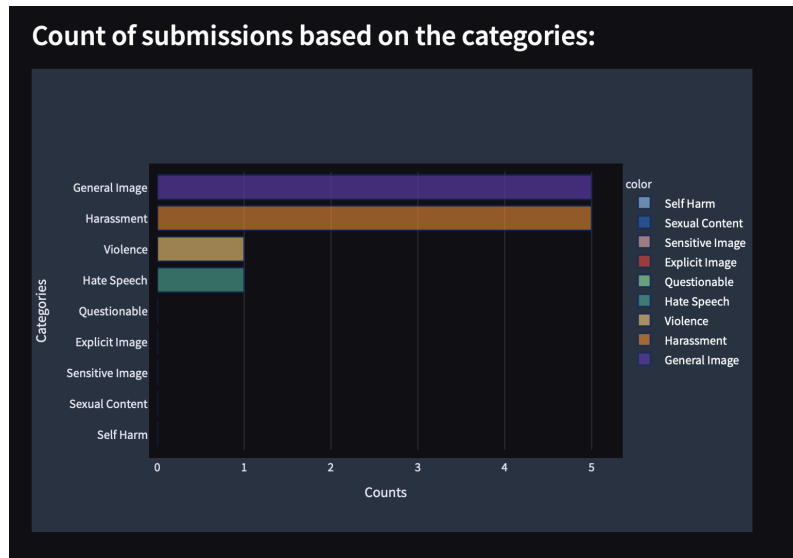
Business users can sign up with SafeFeed and give it moderator permissions for their subreddits. Once connected, SafeFeed's Mage AI system automatically moderates content to keep the community safe and inclusive.

The moderation dashboard shows users important information about their subreddit, like the total number of posts and comments, their sentiment (positive, negative, or neutral),

how many were flagged or removed, the types of violations, and repeat offenders. This helps users manage their communities better.

Users can also see which posts and comments were removed and why, including which policies were violated. This helps users understand the moderation decisions better.





Custom Policies

SafeFeed knows that every online community has different needs. Users can define their own custom policies for their subreddit through a special interface. These custom policies are saved as JSON files and automatically added to the OpenAI Assistant API. This means that moderation decisions are based on the most up-to-date policies for each subreddit.

Reddit Policies Customizer

This tool allows you to edit existing policies or create custom policy files for a specific subreddit.

Select your Subreddit:

Subreddit list

SafeFeed

Fetch policies

Select a policy to edit:

custom

Edit the selected policy:

```
{'movie_policy': "In this community, we do not tolerate negative discussions about movies. Hate speech, derogatory comments, and personal attacks are strictly prohibited. Instead, let's focus on positive aspects and respectful critiques. We believe in creating a welcoming space where members can share their love for cinema without spreading negativity."}
```

Update and Create Custom Policy File

Policy Guide Chatbot

SafeFeed has a chatbot that uses the OpenAI Assistant API to help users understand the platform's policies. Users can ask questions in natural language, and the chatbot will provide clear and relevant information from the policy documents. This chatbot makes it easier for users to understand why certain content was moderated, and promotes a more open and collaborative community.

Select your Subreddit:

Subreddit list

Default policies

Type your message:

can you talk about the harassment violations in the policies?

Send

Conversation

You

can you talk about the harassment violations in the policies?

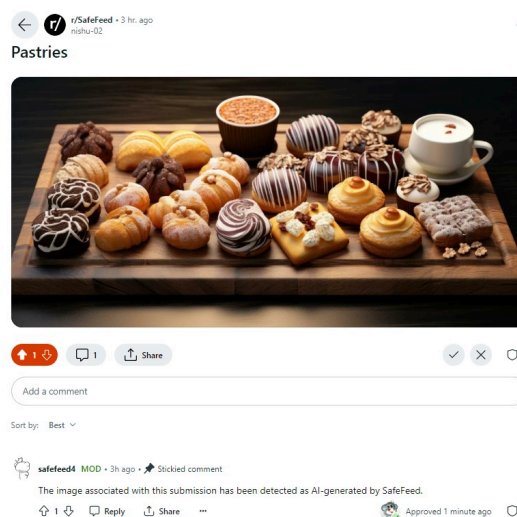
Assistant

The policies strictly prohibit harassment, bullying, and threats of violence on the platform. Here are key points on harassment violations:

- Protection Against Harassment**: Reddit defines harassment as behavior that works to shut someone out of the conversation through intimidation or abuse, whether in public content like posts or comments, or in private messages/chat.
- Types of Prohibited Behavior**: This includes menacing someone, directing abuse at a person or group, following them around the site, and encouraging others to engage in these actions.
- Context of Behavior**: The context and pattern of behavior are considered when evaluating if an action constitutes harassment. Being annoying or downvoting is not considered harassment, but intimidating or abusive behavior is.

AI-Generated Image Detection

SafeFeed also uses Sight Engine to identify if an image uploaded to a post is AI-generated. If an AI-generated image is detected, SafeFeed provides context to the post, indicating that the associated image is AI-generated. This additional context helps users make informed decisions about the content they engage with.





Conclusion

SafeFeed is an innovative tool for content moderation that uses advanced AI technology to help social media platforms maintain safe and welcoming spaces for users while supporting open conversation. It combines powerful language models and image analysis techniques to understand and moderate content effectively.

By combining Reddit and using OpenAI's technologies, SafeFeed shows how AI can tackle difficult problems like content moderation. It makes moderation smarter by understanding the context of discussions, ensuring that online spaces are both safe and open for everyone.