A dataset to evaluate content moderation on Reddit

Ethan Bai
University of Washington
ebai2022@cs.washington.edu

Sehaj Dhillon
University of Washington
sehajd@cs.washington.edu

Elizabeth Wang
University of Washington
uwewang@cs.washington.edu

Abstract

As vision-language models (VLMs) like GPT-4V and Gemini become increasingly powerful, their deployment in subjective applications such as content moderation raises new challenges. Existing benchmarks, such as the Hateful Memes dataset [1], are limited in scope and fail to reflect the diversity of real-world moderation contexts. In this project, we propose a new benchmark that evaluates the ability of VLMs to predict content moderation decisions on Reddit, a platform with decentralized and community-specific moderation policies. We develop a web scraping pipeline that continuously collects multimodal Reddit posts (text and images) across a range of subreddits, identifying whether content has been removed or flagged. Using this dataset, we evaluate the performance of state-of-the-art VLMs in predicting moderation actions and analyzing their alignment with subreddit-specific norms. Our benchmark provides a framework for assessing VLM behavior in nuanced, subjective classification tasks and lays the groundwork for more responsible deployment of AI in online content governance.

1. Introduction

The growing deployment of vision-language models (VLMs) such as GPT-4V, Gemini, and Llava in socially sensitive domains has raised important questions about their reliability, fairness, and alignment. This is especially important in community applications like content moderation where social context tends to vastly change what is acceptable forms of speech. These models are now capable of interpreting complex multimodal inputs, but there is little understanding of how well they handle subjective decisions rooted in community norms and evolving cultural contexts. While benchmarks like the Hateful Memes dataset [1] have enabled progress in multimodal hate speech detection, they fall short in capturing the diversity and ambiguity of real-world moderation environments.

Platforms like Reddit offer a unique opportunity to study subjective content governance at scale. With thousands of independently moderated subreddits, Reddit reflects a spectrum of content policies ranging from strict academic discourse to unhinged humor. Crucially, moderation actions (e.g., post removals, user warnings, content flags) are publicly visible, making Reddit a rich, underutilized resource for benchmarking VLM performance in real-world moderation tasks.

We will create a benchmark specifically for evaluating multimodal content moderation models using Reddit data. We design a continuous scraping pipeline to collect image-text posts across subreddits with varying moderation policies, tracking which posts are removed or flagged and mapping them to relevant community rules. Our dataset will enable evaluation of VLMs on tasks such as predicting whether or not a certain Reddit post should be removed, and by extension, which community guideline is likely being violated. By assessing how well models align with community decisions, our work highlights key strengths and failure modes of existing VLMs in subjective classification tasks.

2. Related Work

Recent efforts in multimodal moderation focus on datasets like Hateful Memes, which combine text and image inputs for hate speech classification. While influential, such datasets are limited to predefined content categories and platform-specific norms, lacking the complexity of real-world, user-generated moderation dynamics. Currently, no standard benchmark to assess how well these models align with community-driven moderation policies.

3. Approach

We propose building a moderation-labeled multimodal dataset by continuously scraping Reddit posts from a curated set of diverse subreddits. Each sample includes:

- Post metadata (title, body text, media)
- Community context (subreddit name, rules, post flair)
- Moderation signals (removal status, mod comments, [deleted] flags, timestamps)

To support scalable and reproducible data collection, we implement an asynchronous scraper loop that continuously polls Reddit's API and archives snapshots of post states. Content that is later removed or flagged will be labeled as moderated, while surviving posts serve as negative examples.

Each post is treated as a multimodal moderation prediction task: given a post's image and text, can a model predict if it will be moderated in its respective subreddit? We will evaluate state-of-the-art VLMs (e.g., GPT-4V, Gemini, Llava) using natural language prompts that simulate moderation judgment, optionally incorporating community rules as context.

4. Planned Experiments

We plan the following experiments:

- Measure how well each model predicts moderation decisions across subreddits, particularly focusing on precision/recall and reducing number of false positives
- Compare model predictions on image inputs versus text inputs versus images and text inputs to see where models perform well
- Evaluate a non-expert human performance versus the performance of each model to see whether it is safe to deploy models in the real world

5. Expected Challenges

We expect to encounter several technical challenges in developing this benchmark. First, Reddit's API rate limits and terms of service will constrain how often and how much data we can collect. In addition, ethical considerations such as anonymizing personally identifiable information and handling sensitive content categories must be carefully addressed to ensure responsible data collection. Second, subreddits have different and often ambiguous moderation policies. This makes it challenging to create a standardized set of labels that can be applied across all subreddits as what is considered offensive in one community may be perfectly acceptable in another. Third, moderation decisions often rely on how nuanced interactions between textual and visual content, which can introduce label noise when the dataset lacks complete post context. Lastly, building a fault-tolerant, continuous scraping pipeline capable of operating over extended periods of time will require significant engineering challenges to prevent any data loss.

6. Expected Outcomes

We expect to deliver the first benchmark dataset for evaluating multimodal content moderation models using Reddit, capturing real-world, community-driven decisions across diverse subreddits. This dataset will enable reproducible and scalable evaluation of VLMs like GPT-4V, Gemini, and Llava on subjective moderation tasks involving both text and image inputs. Our benchmark will provide a diverse and robust set of tests for comparing VLM behavior across domains with varying moderation norms, offer insights into how models currently handle contextual ambiguity, and serve as a foundation for future work in AI alignment, fairness, and interpretability in subjective classification tasks. Ultimately, this work aims to bridge the gap between model capabilities and responsible real-world deployment, offering tools for auditing AI systems tasked with socially sensitive judgments.

References

[1] Douwe Kiela, Hamed Firooz, Vedanuj Mohan, Amanpreet Goswami, Anurag Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2020. 1