Research Design Review

Karl-Johan Westhoff
email kjwesthoff@berkeley.edu

UC Berkeley School of Information
MIDS Course 201 Spring 2025, Section 6 (Brooks Ambrose)

Internal Memo, Content Moderation

Looking at content moderation beyond our own company, there are many organizations with similar challenges to ours. The purpose of content moderation is to ensure that the content complies with community guidelines. The guidelines can be grouped into sections based on the rules specified by each part:

(A) Societal rules: Laws, standards and regulations
(B) Malicious actors: Spam, unwanted commercial and political soliciting, forum hijacking, malware distribution
(C) Social rules: Inappropriate content, Hate speech, Racism, Homophobia
(D) Rules for our company: How do we want people to engage with each other, the debate 'tone', and handling influence, for example from competitors.

Each part can be addressed individually and processed in parallel, and there are various solutions for handling content moderation both in-sourced and out-sourced, AWS for example offers a pay-per-post solution[1]. The aim of this memo is to review the technologies that are available for content moderation and how we should measure how well it works.

1 Potential solutions

1.1 Automation, Machine Learning and AI

An overview of Machine Learning (ML) for content moderation was done by Chowdhury in [2]. Two automated techniques are used in the industry:

- Matching models: Where content is compared (with various degree of similarity) against a database of known unwanted content. This is best for things that are clearly identifiable such as features present in pictures, specific words and language, plagiarism etc.

- Predicting models: Predicts the likelihood of a text or image violating a set of rules defined by the categories above. For example analyses a text to identify if the content violates a law.

Matching models are effective for raw filtering against well defined rules and especially applicable to (A) and (B) above. Predicting models provide a probability that ′something′ is in violation of a rule and includes context rather than just looking at a ′bag of words′.

Features of ML and AI based solutions

- Can scale quickly as traffic increases and decreases on our forums
- Machine Learning based content moderation has historically been prone to over filtering (false positives), with a hands-off approach, operators have erred towards "better catch too many than too few". However, this is a developing field and performance is expected to increase
- Using more advanced methods beyond ML such as Large Language Models (LLM's) on every forum post and thread is currently too expensive. However, the cost is expected to decrease

1.2 Human Moderators

Content moderation using people trained to precisely evaluate the content.

Features, Manual Moderation

- Flexible, high quality evaluation capturing nuances
- Immediate reaction to and prediction of new trends in society
- Low initial cost, high operating cost (requires qualified manpower)
- Scaling with forum traffic is slow, cost increases with increased traffic
- Large cost impact in case of the forum being ′attacked′ with a flood of malicious posts

1.3 Hybrid Approach

The hybrid approach seems to be the industry norm, with efforts to increase the automated part. Sircar et al. looked at the company ′X″s transparency report in [3]. The largest players in social media are trying to increase the reliance on AI, with some unexpected side effects: AI seems not yet to be ready to do the work alone. "AI systems struggle to identify the full spectrum of harmful behaviors"[3], training data are skewed to ′western′ societies and misses hate speech from other cultures and does not understand features such as sarcasm and ′secret language′[1].

[1] Like kids inventing their own language and expressions

Features, Hybrid Moderation

- Cost-efficient compared to a purely human approach
- Can react quickly to new language, words and trends
- Lets AI/ML do the parts it is good at, start with from 'Matching Models' and expand from there
- More complex to implement and requires constant adjustment as automated processes can do more
- Operators can transition from doing the content moderation itself to using domain knowledge for improving the automation

2 Evaluation

2.1 Trial Stage Evaluation

The purpose is to apply measurements of how good the proposed content moderation approaches above work prior to committing fully to a solution. In this case we do have a set of labeled data from our current content moderation which can serve as training and test data for evaluation.

Trial Test Setup

- Run each solution with the test data
- Evaluate each solution, especially for false negatives and false positives
- Compare and review the solutions including cost of implementation and operation
  - The possible future improvements to automated solutions should be evaluated and included
  - The evaluation should be done as a pilot project as part of the research design

Trial Metrics Classical machine learning metrics[4]ch.2:

- Precision, How good did we predict, false positives
- Recall, How good did we predict false negatives
- F1 Score, combination of precision and recall

2.2 Evaluation Post-Deployment

We can use the same Machine Learning metrics as for the trials by sampling data for bench marking as we go along. However, the criteria for what we consider inappropriate may change over time[2]. This can be evaluated using surveys of customer satisfaction and usages statistics. Cost vs. efficiency of our content moderation should be evaluated periodically as new technology evolves.

[2] especially for (D) above regarding company defined rules

In-Operation Metrics

- Periodical sampling and evaluation of ML
  - Precision
  - Recall
  - F1 Score
- Customer satisfaction surveys, will indirectly compare to the rest of the industry
- Usage statistics, tied into business goals and performance metrics
- Periodic evaluation of cost

Conclusion

Three solutions for content moderation were proposed and methods for evaluation during design and after implementation were evaluated. The solutions represent the current industry standards and near future improvements. The evaluation metrics are based on classical methods and comparison to industry standards and market competitors.

References

[1] AWS. *Protect your users, brand, and budget with AI-powered content moderation*. `https://d1.awsstatic.com/psc-digital/2022/gc-400/ml-leading-use-case-ebook/204863_AWS_ML_Content_Moderation_eBook_Final.pdf`. Accessed: 2025-02-21.

[2] Nafia Chowdhury. *Automated Content Moderation: A Primer*. `https://cyber.fsi.stanford.edu/news/automated-content-moderation-primer`. Accessed: 2025-02-21.

[3] Anisha Sircar. *X's Latest Content Findings Reveal Troubling Trends In AI Moderation*. `https://www.forbes.com/sites/anishasircar/2024/10/18/xs-latest-content-findings-reveal-troubling-trends-in-ai-moderation/`. Accessed: 2025-02-21.

[4] David Freeman Clarence Chio. *Machine Learning and Security*. Accessed: 2025-02-21.

Notes on this assignment: Word count of the body text is approx. 970 words, A three page 'lorem ipsum' document with 5 paragraphs and 1in margins would be up to 950 words, I thereby argue that the assignment fulfills the assignment length requirement by being equivalent to a Three page document.
I fed the assignment to ChatGPT, it gave me some ideas to the overall structure of the assignment, and the 3 types of content moderation (although it was somewhat obvious) you can find my 'conversation' here: `https://chatgpt.com/share/67bcb191-e528-8008-b806-3b2f8c463a0f`