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

Distribution list: G.B.O.A.T, C.D.S.A, B.I.M.P.S[1]

Overview

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 be subdivided into parts:

(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 both in-sourced and out-sourced f.ex. AWS[1] offers a pay-per-post solution. The aim of this memo is to review the technologies that are available for content moderation and how we should measure effect. Who does it on which part is not part of the scope.

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:

[1] Big Important Manager Persons

- 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 featured 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 can be considered raw filtering and especially applicable top (A) and (B) above. Predicting models is where the current development is focused.

Features

- 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

1.2 Human Moderators

Content moderation using people trained to precisely evaluate the content.

Features

- Flexible evaluation capturing nuances
- Immediate reaction to new things in society
- Low initial cost, but requires more manpower
- Scaling with forum traffic is slow, cost increases with increased traffic

1.3 Hybrid Approach

Combining the approaches.

Features

- Cost-efficient compared to a purely human approach
- Can react quickly to new language, words and trends f.ex. in the hate speech landscape
- 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

- Can transition from doing the content moderation itself to using domain knowledge for improving the automation

The hybrid approach seems to be the industry norm, Sircar et al. looked at the company 'X"s transparency report in [3]. The largest players in social media are increasing reliance on AI with some unexpected side effects, AI seems not yet 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 or over flags benign entries as hateful and does not understand sarcasm and 'secret language'[2].

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 of the effect of each solution.

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 especially for (D) above, 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] Like kids inventing their own language and expressions

In Operation Metrics

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

3 Conclusion

Three solutions for content moderation was and methods for evaluation during design and after implementation were proposed. 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 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.