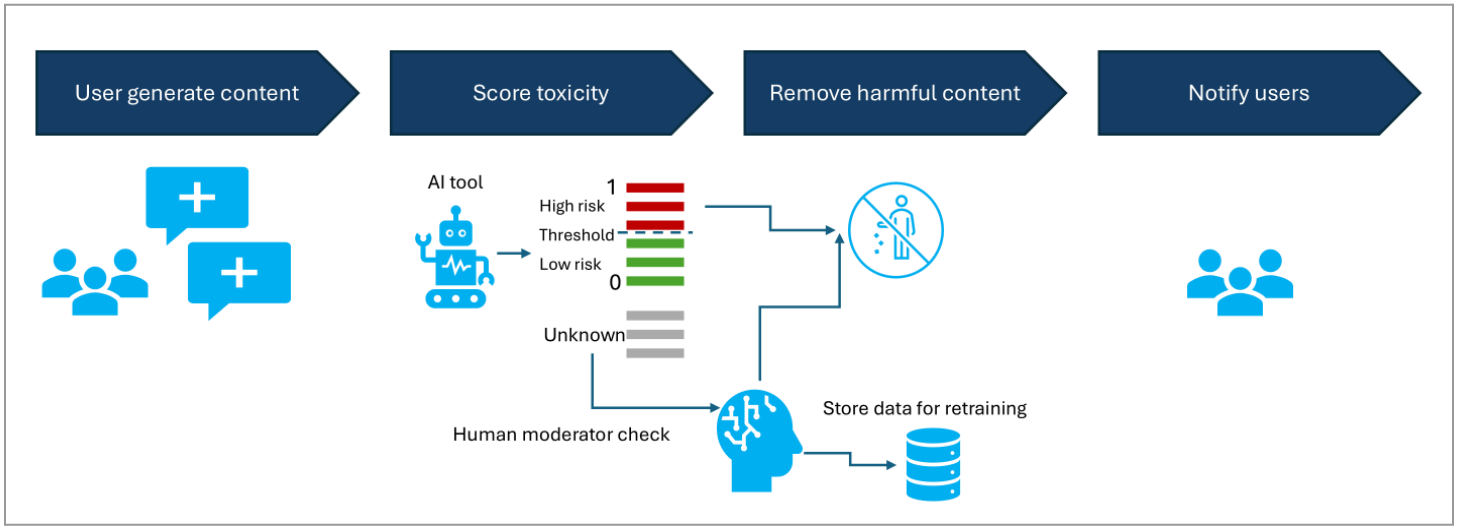
Yoko Morishita, DASCI 201, Section 02
October 10, 2024

Unit 08 Assignment

1. **Context & Assumption:** The company provides an online news publication platform, and we need to monitor and apply guidelines to user-generated content (ie. posts via the comments section of the website) by identifying and removing spam, offensive language, or inappropriate content. I assume the "user-generated content" includes only text comments.

2. **Potential Solutions:** The image below shows what the overall process looks like when we follow the industry trend. The basic concept is to combine AI-based automated moderation and human moderators. AI will filter content into "High risk", "Low risk" and "Unknown Risk" and human moderators will review the decisions for Unknown cases. Once flagged content is added to the queue, it will be removed automatically, and the user who made the content will be notified.



2.1. For AI-based moderation, multiple options exist:

	Description	Pros	Cons
Existing tool	Connect existing tools to our flow through API (ex. Perspective API)	Low-cost Quick/simple to deploy Limited maintenance required Less biased (trusted model)	May not fit our requirements

Customized tool	Use existing AI model (eg. Chat GPT content moderator) and train it	Higher accuracy can be expected, especially in nuanced cases	May not fit our requirements Tuning can take time and cost Need maintenance
Newly developed tool	Develop our own model from scratch	Customizable Optimizable Can fit our requirements Fully transparent	Tuning can take time and cost May not be scalable Need maintenance

3. Evaluation (before implementation):.

3.1. Design:

3.1.1. **Objective**: Evaluate the performance of each tool in identifying harmful content (true positives) and correctly allowing appropriate content (true negatives) with minimal errors (false positives/negatives), and eventually choose one solution to implement.

3.1.2. **Assumptions:** Categories that violate the guidelines are hate speech, threats, and personal attacks. What we need to discuss before starting project:

3.1.2.1. Scope of the content (eg. any future plan to expand beyond texts?)

3.1.2.2. Definition of each category (eg. what is the threat?)

3.1.2.3. Toxicity score (0 -1 scale which shows how likely the content will be regarded as harmful content and threshold between High risk and Low risk)

3.1.2.4. Language and Region

3.1.2.5. Target metrics (Rule of thumb is 85%-90% for all metrics)

3.1.2.6. Duration and milestones of the project

3.1.2.7. Target sample data volume

3.1.3. Metrics:

3.1.3.1. **Precision**: The proportion of true positives among flagged content (true positive + false positive).

3.1.3.2. **Recall**: The proportion of true positives among actually true content (false negative + true positive).

3.1.3.3. **Accuracy**: The overall percentage of correct classifications ((true positive + true negative) / total).

3.1.3.4. **F1score**: The harmonic mean of precision and recall: (precision*recall) * 2 / (precision + recall)

3.1.3.5. **Processing Time**: The speed of content moderation for each tool.

3.1.4. **Dataset:** Historical data. It has to be high volume, diverse, representative (including such as gray or culturally specific content), and accurately labeled. If needed, 70% will be used to train and tune the model, and the remaining will be used for evaluation.

3.2. Evaluation of test analysis

3.2.1. **Performance analysis**: Compare the results of the tools based on metrics.

3.2.2. **Cost:** Evaluate costs per tool, especially with the volume of content the company will need to moderate.

3.2.3. **Scalability and Elasticity:** Consider how each tool performs with larger volumes of data and more complex content. Also, consider if we want to expand the category in the future.

4. **Evaluation (after implementation):** We should have a pilot program until we have enough sample data, which allows us to monitor, report monthly and make judgement if new solution can improve the status quo.

4.1. A/B Testing:

4.1.1. A pilot program with two different groups: 1. the new tool is active, and 2. the old moderation method is used.

4.1.2. Comparison: Compare the user engagement (eg. comments posted), flagged content, and user-reported satisfaction (e.g., how users react to flagged content or removed posts) in each group.

4.2. Monitoring:

4.2.1. Live data performance: Using the metrics, monitor the tool performance.

4.2.2. Random check: AI to randomly pick content and human moderators (internal moderators or a small group of users) should be in loop to see if the tool judgement is correct.

4.2.3. **Human feedback:** Establish a feedback loop from users and human moderators.

4.3. Tuning: Assess how much effort would be needed to keep using the tool.

4.3.1. **Training and Retraining**: Periodically retrain the AI models using new data to improve performance on evolving content trends and language, and limit bias.

4.3.2. **Threshold Adjustment**: Over time, adjustment the toxicity thresholds may needed (eg. due to changes in people's sensitivity).

4.4. Cost: Conduct cost-performance analysis, especially considering how much it would cost to implement new solution (tool cost, human moderator and maintenance)

4.5. Risk and other considerations: Consider what sort of other positive/negative effect that solution can bring: eg. By reducing human moderator's exposure to harmful content, new solution can bring positive psychological effect

4.6. Management setup:

4.6.1. Create dashboard with key metrics

4.6.2. Set up meeting cadence, R&R and decision making process