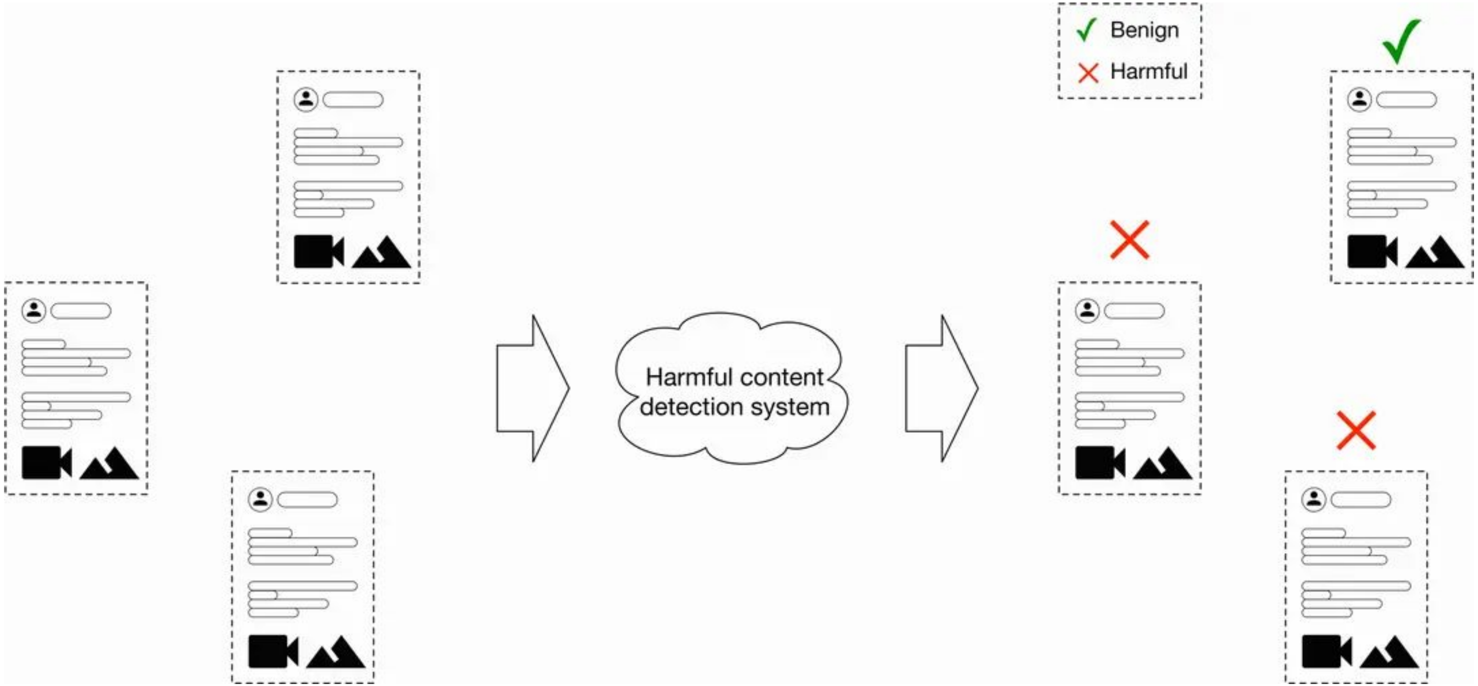
Harmful Content Detection

By 夏之未至 Jul. 27, 2025

Outline

- Clarifying Requirements
- Frame the Problem as an ML Task
- Data Preparation
- Model Development
- Evaluation
- Serving


Benign
Harmful
Harmful content detection system

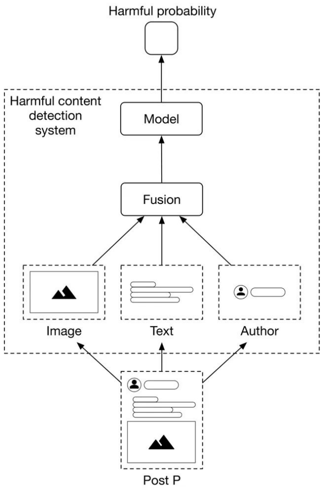
Clarifying Requirements

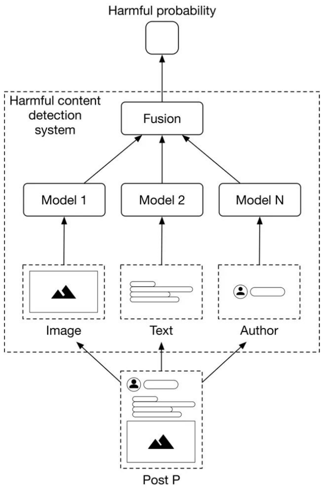
- Definition of harmful content
  - Harmful content: Posts that contain violence, nudity, self-harm, hate speech, etc. (*)
    - To be discussed: violence, nudity, hate speech
    - Too complex, out of range: misinformation
  - Bad acts/bad actors: Fake accounts, spam, phishing, organized unethical activities, and other unsafe behaviors.
- Input
  - content format
    - text, image, video or combination
  - Language
  - Label Data availability: 500 million posts/ day, annotate 10,000/day
    - Annotation: expensive, time consuming
    - User reports
- Output
  - Classification & explanation
- Latency
  - Real-time: detect and block immediately
  - Batch: detect offline hourly/daily

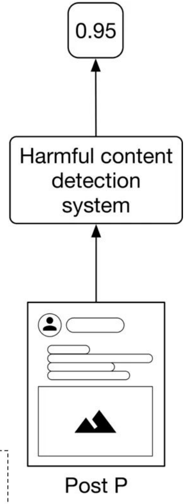
Frame the Problem as an ML Task

- Predicts probability → classification

- Multiple format of input → multimodalities
  - Fusion of different modalities
    - Early fusion: left
    - Late fusion: right







Frame the Problem as an ML Task

Aspect	Late Fusion	Early Fusion √
Model Independence	train/evaluate/ improve independently	Single model, all modalities trained together
Data Requirement	separate training data for each modality (time-consuming, expensive)	Unified training data needed for the joint model
Detection of Harmful Combinations	May fail: If each modality is benign, combined output can miss harmful interactions	Can detect: Model considers all modalities jointly, so can capture harmful combinations
Training Complexity	Simpler	More complex; model learn relationships between modalities
Data Sufficiency	Works even with limited data	Large data to learn complex cross-modal relationships
When Preferred	When modalities are independent, data is scarce, or separate improvement is needed	When harmfulness arises from cross-modal interaction and plenty of data is available
Example	Individual meme image/text models miss harmful memes	Detect harmful memes where image and text are benign alone but harmful together
Preferred in This Case	Not recommended due to inability to capture cross-modal harm	Recommended, since ample data allows model to learn complex, harmful cross-modal patterns

Frame the Problem — right ML category

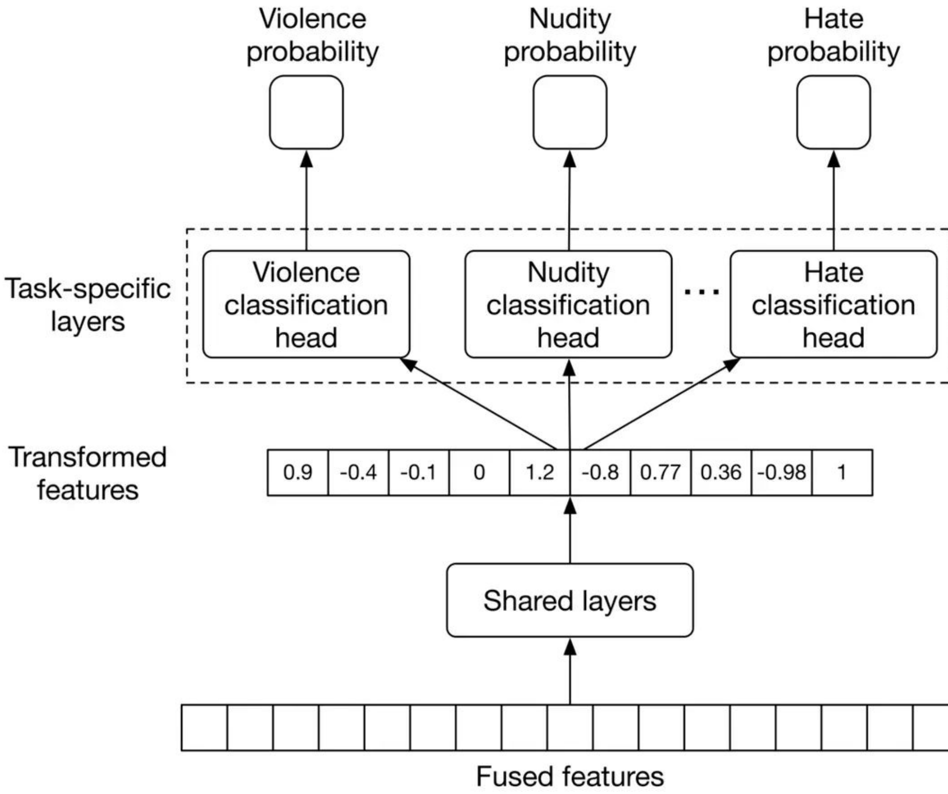
Approach	Pros	Cons
Single Binary Classifier	Simple; easy to implement	Cannot explain class of harm; cannot track class-wise errors; poor for actionable feedback
One Binary Classifier per Class	Explains which class caused removal; can monitor/improve models independently	Training/maintaining many models is costly and time-consuming
Multi-label Classifier	Only one model to train/maintain; predicts all classes at once	Shared features may not fit all classes equally; may need different input transforms
Multi-task Classifier	Shares learning across classes; efficient with computation and data	More complex architecture; may require more tuning and careful task balancing

Multi-task classifier overview

- Shared layers: transform input features into new ones
- Task-specific layers

Pros

- Efficient Training & Maintenance as a single model
- Shared Feature Transformation
- Data Efficiency: good for limited data



Data Preparation

- Users

ID	Username	Age	Gender	City	Country	Email

- Posts

Post ID	Author ID	On-device	Timestamp	Textual content	Images or videos	Links
1	1	73.93.220.240	1658469431	Today, I am starting my diet.	http: //cdn.mysite.com/u1.jpg	-

- User posts interactions

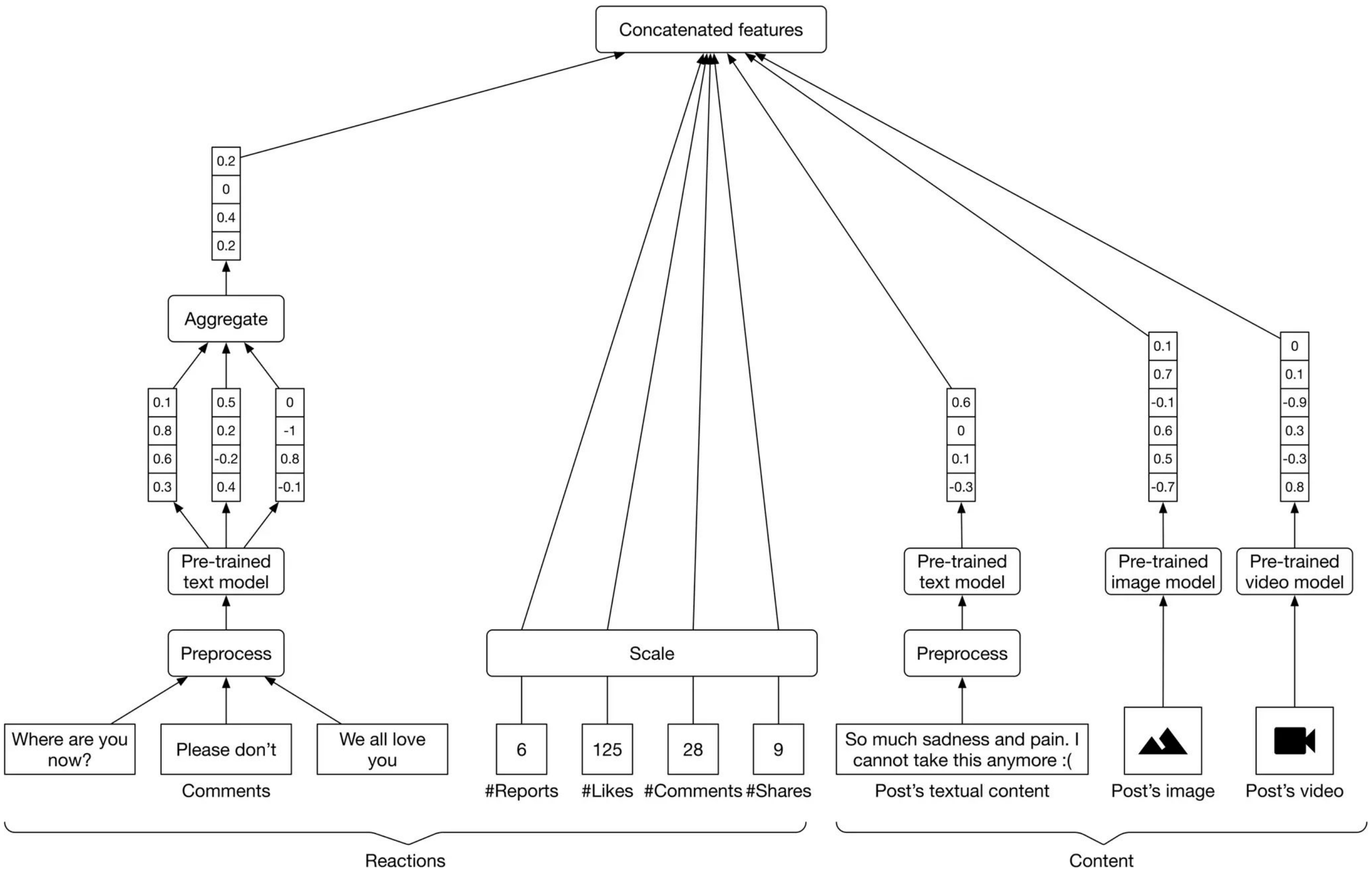
User ID	Post ID	Interaction type	Interaction value	Timestamp
11	6	Impression	-	1658450539
4	20	Like	-	1658451341

Feature engineering

- Posts
    - Textual content
        - Text preprocessing
        - Vectorization
            - Fast & easy: BoW or TF-IDF but no semantic info
            - Pretrained LLM: DistilmBERT, an efficient variant of BERT to handle the latency and multilingual words embeddings
    - Image or video
        - Preprocessing: decode, resize, and normalize the data.
        - Feature extraction: CLIP or SimCLR, VideoMoCo: unstructured data→ feature vector

Feature engineering

- User reactions to the post
    - The number of likes, shares, comments, and reports: scale these numerical values to speed up convergence during model training.
    - Comments: pre-trained model → word embeddings → aggregate(eg: avg) to get the embedding for the comment

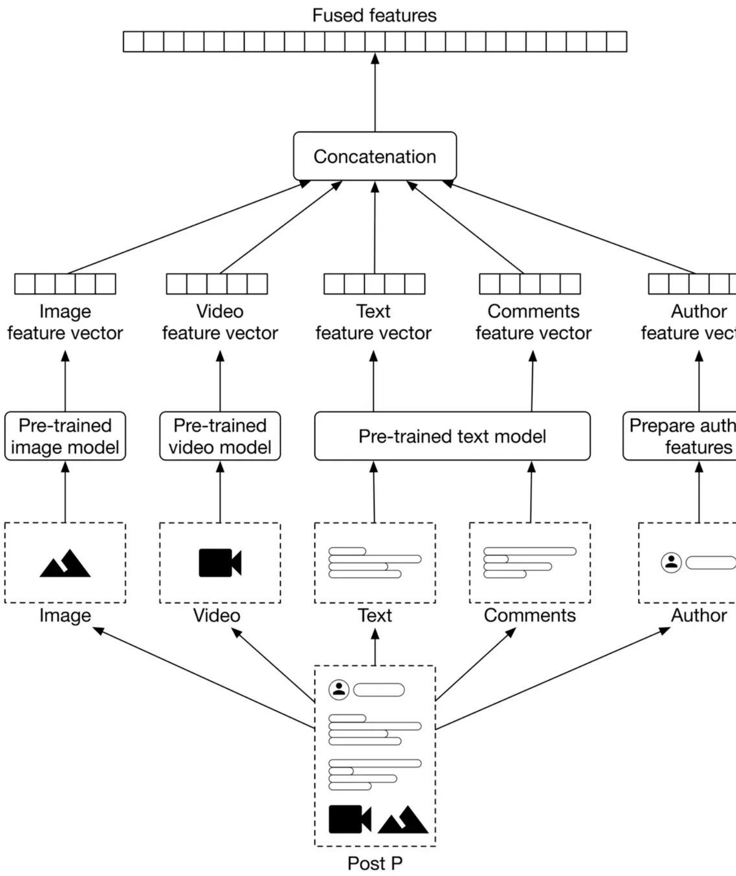

Concatenated features
0.2
0
0.4
0.2
Aggregate
0.1
0.8
0.6
0.3
0.5
0.2
-0.2
0.4
0
-1
0.8
-0.1
Pre-trained text model
Preprocess
Where are you now?
Please don't
We all love you
Comments
Scale
6
125
28
9
#Reports
#Likes
#Comments
#Shares
0.6
0
0.1
-0.3
Pre-trained text model
Preprocess
So much sadness and pain. I cannot take this anymore :(
Post's textual content
0.1
0.7
-0.1
0.6
0.5
-0.7
Pre-trained image model
Post's image
0
0.1
-0.9
0.3
-0.3
0.8
Pre-trained video model
Post's video
Reactions
Content

Feature engineering

- Author
  - Author's violation history
    - Number of violations
    - Total user reports
    - Profane words rate: profane words are predefined.
  - Author's demographics
    - Age: one of most important predictive features
    - Gender: 1-hot
    - City and country: embedding layer, not 1-hot since sparse.
  - Account information
    - Number of followers and followings
    - Account age

Feature engineering

- Contextual information
  - Time of day:
    - bucket:
      - morning, noon, afternoon, evening or night
    - 1-hot embedding
  - Device: 1-hot embedding

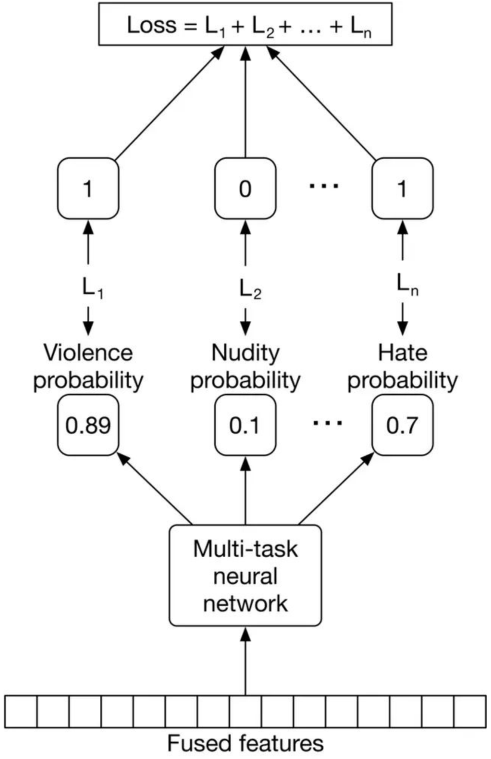


Model Development — Constructing the dataset

Labeling Method	Description	Pros	Cons	Usage
Hand Labeling	Human contractors manually label posts	Accurate labels	Expensive, time-consuming	Evaluation data
Natural Labeling	Uses user reports for automatic labeling	Fast labeling	Noisier, less accurate	Training data

Model Development – loss function selection

- Binary classification loss for each task
  - Eg: cross entropy
- Overall task: combining task-specific losses
  - Eg: sum of each loss
- Challenge
  - Learning speed difference among modalities and one dominates
  - Resolution
    - Gradient blending
    - Focal loss



Evaluation

- Offline
  - PR-RoC trade-off between precision and recall (better for imbalanced data)
  - ROC curve.
- Online
  - Prevalence. This metric measures the ratio of harmful posts which we didn't prevent and all posts on the platform.

$$\text{Prevalence} = \frac{\text{Number of harmful posts we didn't prevent}}{\text{Total number of posts on the platform}}$$

  - Harmful impressions.
  - Valid appeals

$$\text{Appeals} = \frac{\text{Number of reversed appeals}}{\text{Number of harmful posts detected by the system}}$$

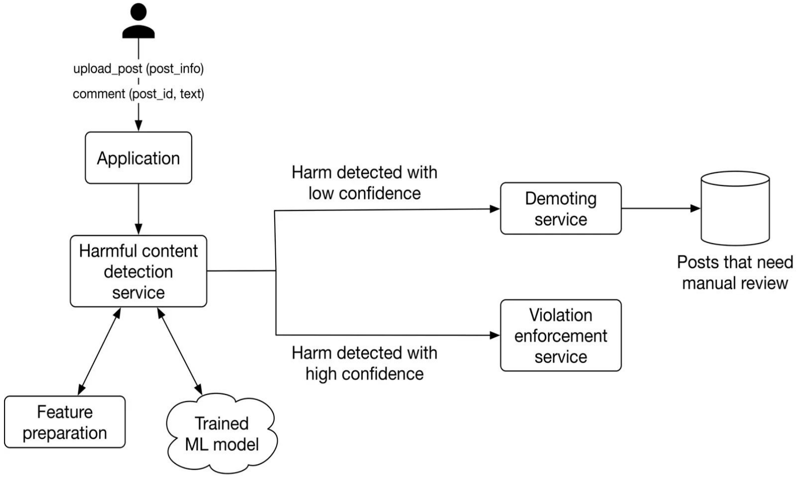
  - Proactive rate

$$\text{Proactive rate} = \frac{\text{Number of harmful posts detected by the system}}{\text{Number of harmful posts detected by the system} + \text{reported by users}}$$

  - User reports per harmful class.

Serving

- Harmful content detection service
  - predicts the probability of harm
- Violation enforcement service
  - immediately takes down a high confidence harmful post by detection service
- Demoting service
  - Demote the low confidence harmful post
  - Store it in the storage for manual review



Additional topics to talk

- Handle **biases** introduced by human labeling
- Adapt the system to detect **trending** harmful classes (e.g., Covid-19, elections)
- How to build a harmful content detection system that leverages **temporal information** such as users' sequence of actions.
- How to **effectively select post samples** for human review
- How to **detect authentic and fake accounts**
- How to deal with **borderline contents** [25], i.e., types of content that are not prohibited by guidelines, but come close to the red lines drawn by those policies.
- How to make the harmful content detection system **efficient**, so we can deploy it **on-device**
- How to substitute Transformer-based architectures with linear Transformers to create a more efficient system.