机器学习系统设计题：Unsafe content 检测

MLE 算法指北

2025 年 5 月 22 日

1 问题定义与背景

在内容社区平台（如 tiktok、ins、Pinterest、Reddit、Twitter 等）中，用户发帖可能涉及各种 **不安全内容（NSFW，Not Safe For Work）**，包括：

- 色情、裸露、性暗示；
- 暴力、血腥、自残；
- 仇恨言论、骚扰、歧视；
- 虚假内容、诈骗信息；
- 灰产、毒品、枪械相关。

这些内容不仅影响用户体验，也可能引发法律/合规风险。如何通过机器学习自动检测并拦截 unsafe content，是一个典型的多模态机器学习系统问题。

2 系统输入与模态分析

用户发帖内容通常包含如下模态：

- **文本信息**：标题、描述、评论、OCR 识别出的图中文字；
- **图像信息**：主图、预览图、头像等；
- **上下文信息**：用户行为、地理位置、链接域名；
- **视频/音频（可选）**：通过抽帧 + 语音识别进行转换。

3 整体系统架构

3.1 核心模块

(1) **预处理模块**：包括分词、OCR 提取、ASR 转写、图像压缩等；

(2) **文本检测模型**：BERT / RoBERTa + 多标签分类器，检测色情词、暴力、仇恨；

(3) **图像检测模型**：使用 RESNET, VIT, CLIP, YOLO 等模型识别裸露、血腥等；

(4) **多模态融合**：拼接文本向量和图像向量，通过 MLP/Transformer 做联合分类；

(5) **输出判定模块**：基于规则 + 模型分数打标签，返回结果如：

```
{
  is_nsfw: true,
  labels: ["nudity", "violence"],
  confidence: 0.94
}
```

4 模型设计与训练

4.1 文本模型

4.1.1 输入结构与建模框架

模型输入包括原始文本（标题、描述等）与 OCR 提取出的图片中文字，采用拼接方式输入：

$$\text{Input} = \texttt{[CLS]}\ \text{Post Text}\ \texttt{[SEP]}\ \text{OCR Text}\ \texttt{[SEP]}$$

模型结构如下：

$$x = \text{BERT}(T), \quad h_{\text{CLS}} = x_0, \quad \hat{y} = \sigma(W h_{\text{CLS}} + b)$$

根据任务需求可为二分类或多标签分类输出。对于多任务（如成人/暴力/仇恨），可使用多个独立分类头：

$$\hat{y}^{(i)} = \sigma(W_i h_{\text{CLS}} + b_i), \quad \mathcal{L} = \sum_i \lambda_i \cdot \mathcal{L}_i$$

4.1.2 文本增强策略

- **拼写扰动增强**：对 “luo 体”、“s3x”、“骚　擾” 等变体词构造映射，并在训练时随机替换 token，以提高模型对规避攻击的鲁棒性。
- **多任务联合建模**：联合 toxic speech、hate speech、offensive 语言检测任务，通过共享 encoder 提高泛化能力。
- **OCR 文本融合**：提取图片中文字，与主文本拼接输入，覆盖 “图文擦边” 内容，提高 recall。

4.1.3 训练与优化策略

- **数据增强**：使用 DropWord、Mask token、拼写错乱、句子乱序等策略，提升泛化能力；
- **类别不均衡处理**：使用 class-balanced loss 或 Focal Loss 解决 NSFW 类别稀缺问题；
- **难样本挖掘**：引入 hard negative mining 策略，关注 “伪安全” 样本（如擦边文案）；

4.1.4 策略开发

- 对每个标签设置不同阈值；
- 中置信度样本进入异步复审队列；
- 可加入 attention-based trigger 提取模块，辅助生成审核理由；

4.2 图像模型设计：用于检测 NSFW 视觉内容

为了有效识别图像中的 NSFW 信息，系统需具备多标签分类、区域检测、图文结合等多种识别能力。

4.2.1 主流图像模型选择

当前可选的图像模型包括：

- **图像分类模型**：RESNET、VIT 给检测图像进行分类。识别低质图片内容。
- **敏感区域切割**：使用 YOLO 进行检测和人体部位分割。可返回图像中裸体区域的位置坐标，适合用于 “裸露评分”、“遮挡区域提取” 等场景。
- **CLIP + Prompt**：通过构造自然语言 prompt（如 “a photo containing nudity”）与图像嵌入对比，计算匹配度。具备跨模态、零样本能力，适合扩展到新类型内容（如血腥、猎奇、政治符号等）。

4.2.2 预处理与数据增强策略

为了提升模型的泛化能力与鲁棒性，可对图像进行如下增强：

- **马赛克增强**：模拟用户上传的打码内容，训练模型识别“部分裸露”、“遮挡但可判定”的图像。
- **随机遮挡 + occlusion dropout**：随机在图像上打补丁或遮蔽，增强模型识别部分可见敏感区域的能力。
- **色彩扰动 + 模糊处理**：模拟用户上传的低清晰度图像或滤镜图，提升对压缩攻击、模糊攻击的鲁棒性。
- **人体检测辅助**：使用 OpenPose、YOLO-pose 等模型检测出人体关键点，在检测中引入如下规则：
  - 识别“脸、胸、胯”区域并着重分析该区域纹理；
  - 若图像中出现大比例暴露区域，提升判定概率；

4.2.3 多标签与类型扩展

除二分类（NSFW/SAFE）外，系统通常需支持多种类型标签：

- 成人（Adult）；
- 暴力（Violence）；
- 恐怖/恶心（Gore）；
- 政治/极端符号；
- 疑似灰产引流（如美女 + 二维码）；

推荐使用 sigmoid 多标签结构输出，或采用 CLIP 结构，通过 prompt 构造标签组：

$$\texttt{Prompts} = [\text{“nudity”}, \text{“gore”}, \text{“political symbol”}, \dots]$$

并与图像 embedding 做 cosine similarity 匹配，选出相似度最高类别。多目标训练相关知识可以看之前笔记。

4.2.4 集成策略与系统融合

为提升鲁棒性与覆盖面，推荐集成多个图像模型：

- 图片分类模型用于快速前置筛查；
- 敏感区域识别用于区域定位与分析；
- CLIP 用于扩展检测类目并捕捉潜在 prompt 配对；

可以采用平均加权或投票融合方式，或使用 MLP 学习不同模型输出的融合权重：

$$\hat{y} = \text{MLP}([y_{\text{OpenNSFW}}, y_{\text{CLIP}}, y_{\text{NudeNet}}])$$

系统最终输出可包括：

- 是否 NSFW；
- 多标签类型（色情 / 血腥 / 恐怖 / 政治）；
- 图像置信分数；
- 可选的检测框坐标；

4.3 多模态融合策略：文本与图像联合建模

在实际内容审核任务中，仅依赖图像或文本单一模态可能会导致严重的误判。例如：

- 图像无裸露，但描述具有性暗示；
- 文本无问题，但图像为擦边色情图；
- 图文组合产生歧义，诱导用户跳转非法内容；

因此，需要将图像向量 I 与文本向量 T 融合，构建统一表示向量 h，用于判断内容是否为不安全（NSFW）及其类型。

4.3.1 方法一：向量拼接 + 非线性融合（MLP）

这是最简单、最实用的融合方法：

$$h = \text{MLP}([T; I; T \cdot I])$$

其中：

- T：文本模型（如 BERT）的 `[CLS]` 表示；
- I：图像模型（如 ViT 或 CLIP）的图像全局表示；
- $T \cdot I$：element-wise 乘法（交互信息）；
- MLP：多层感知机（Linear + ReLU + Dropout）；

该结构简单易部署，适合对 NSFW 等二分类/多标签任务快速上线。

4.3.2 方法二：交叉注意力机制（Cross Attention）

该机制用于建模文本和图像之间的 token-level 对齐：

$$A = \text{softmax}(Q_T K_I^\top), \quad H = AV_I$$

- Q_T 来自文本 token；- K_I, V_I 来自图像 patch；- 最终 H 融合图像信息后返回给文本 decoder 或分类器。

可用于对齐视觉区域与敏感描述（如“胸”、“下体”），适合中等复杂场景，如对图文一致性要求高的平台。

4.3.3 方法三：门控机制（Gated Fusion）

为避免某一模态 dominate，采用可学习权重融合策略：

$$g = \sigma(W_1 T + W_2 I + b), \quad h = g \cdot T + (1 - g) \cdot I$$

- g：模态门控向量（$0 \sim 1$），可学习；- 可扩展为多维门控（即对每一维特征学习门控）。
该方法适合模态信噪比差异大的场景（如图像模糊但文本明确）。

4.3.4 方法四：双塔结构 + 后融合

将文本和图像分别通过单独的 encoder 编码，然后在后续使用一个融合层（如 MLP 或注意力）进行联合分类：

$$T = f_{\text{text}}(x_t), \quad I = f_{\text{image}}(x_i), \quad h = \text{FusionLayer}(T, I)$$

可灵活替换 encoder，适合大规模异构模型组合，支持模块化部署与缓存加速。

4.3.5 方法五：CLIP 直接对比（多 prompt matching）

使用 CLIP 结构对图像和文本分别编码，并计算相似度矩阵：

$$\text{score} = \cos(f_{\text{img}}(x_i), f_{\text{text}}(p_k))$$

其中 p_k 为预定义的 NSFW prompt（如 “a photo with nudity”，“an image with violence”）。该方法适用于无标签或标签不完整的场景，具有强扩展性。

5 训练数据构造策略

5.1 数据来源

- 人工审核日志（带标注）；
- 用户举报内容；
- 抓取外部站点内容（色情/暴力论坛）；
- 利用规则引擎进行弱标签标注；
- 模拟生成内容（如 Stable Diffusion 生成裸露图）；

5.2 多轮自训练 + 人审闭环

流程：

1. 初始模型在小规模高质量数据训练；
2. 用模型打伪标签，选出高置信伪样本；
3. 与原数据混合训练；
4. 通过人工审核高不确定性样本；

最终模型使用半监督 + 人审闭环迭代优化。

6 一致性增强与对抗防御

6.1 Consistency Loss

鼓励模型对输入扰动前后保持预测一致：

$$\mathcal{L}_{\text{cons}} = D_{KL}(f(x) \parallel f(T(x)))$$

其中 $T(x)$ 是文本/图像扰动（如 mask、打乱顺序、拼写错误等）。

6.2 对抗防御策略

- Dropout + Embedding Noise；
- 拼写扰动样本增强（针对文本攻击）；
- 图像遮挡模拟 + 模糊增强；

7 部署与防规避机制

- **异步审核队列**：对不确定样本延迟展示、排队人工审核；
- **随机扰动机制**：避免用户过拟合模型决策边界（如换背景逃避）；
- **置信度分级处理**：0.95 以上直接屏蔽，0.5 0.95 进入灰度审核；
- **多版本 ensemble**：提升稳定性与安全性；

8 总结

NSFW 检测是一个多模态、高对抗、动态演化的问题。设计一个高质量的内容审核系统应遵循：

- 多模态协同：文本 + 图像 + OCR + 行为；
- 弱监督放大：规则 → 自训练 → 半监督；
- 鲁棒性增强：一致性 + 对抗训练；
- 可运营性：与人工审核深度结合，建立反馈闭环；

未来工作可以引入：

- Vision-Language 大模型（如 Flamingo, BLIP）；
- 多标签细粒度分类（擦边 vs 明显裸露）；
- 用户层级建模（主动违规者识别）；