# 伪文明评论研究设计

## 一、背景信息（Background）

在线评论中的不文明现象并非都以粗口或直接侮辱呈现；大量“披着礼貌外衣的攻击”（如讽刺、反话、冷嘲）在平台上广泛存在，往往能绕开简单的脏话词典或基于显性攻击特征的模型。  
这类伪文明评论（pseudo-civility）在文本表层呈现礼貌、克制或理性，但在语用层面传达贬低、嘲讽或攻击，其危害在于：更易扩散、更难监管，也更容易被自动化系统“误判为文明”，造成漏检或治理真空。  
现实需求：现有模型对“显性不文明”识别不错，但在“隐性/讽刺性不文明”上表现不稳，导致文明误杀与不文明漏检并存。

## 二、理论框架（Theoretical Framework）

1. 双系统理论（Dual-Process Theory, Kahneman）  
系统1：快速、情绪化、自动化反应（冲动性表达、攻击更集中）。  
系统2：缓慢、理性、受控加工（礼貌、克制、结构更复杂）。  
命题：伪文明是系统1的消极情绪被系统2“语言包装”后的产物——语义更接近文明，情绪/意图更接近不文明。  
  
2. 礼貌理论（Politeness Theory, Brown & Levinson）  
语言服务于“面子”管理：礼貌策略可维护或威胁他人面子。  
命题：伪文明利用“礼貌形式”掩盖“面子威胁”，形成语用层的隐性攻击（mock politeness）。  
  
3. 讽刺/反语语用理论（Sarcasm & Irony）  
讽刺的本质是语义–情绪/语境不一致（semantic–affective incongruity）。  
命题：当“词面积极”与“情感/语境消极”相冲突时，更可能是伪文明。  
  
整合模型（文字版）：  
情绪激活（系统1，负向情绪） → 攻击意图  
文本礼貌包装（系统2，礼貌策略） → 伪文明（表礼貌/里攻击）

## 三、研究方法（Method）

3.1 研究问题与假设  
RQ1：伪文明在语义/情绪特征上与文明、不文明分别有什么系统差异？  
RQ2：加入“语义–情绪不一致”特征后，模型能否更好识别伪文明？  
  
假设：  
H1：伪文明的语义向量更靠近文明，但情绪指标更靠近不文明。  
H2：伪文明具有显著更高的语义–情绪不一致度。  
H3：在引入“礼貌强度 + 不一致度”特征后，模型对伪文明的识别F1提升显著。（代码中现在还没有）  
  
3.2 数据与标注  
语料：沿用现有评论数据（含文本与是否移除标签）。  
三类定义：文明、伪文明、不文明。  
  
3.3 特征工程  
礼貌强度：礼貌词频、缓和语、请求结构等。  
情绪/语调：情感分、符号密度、大写比例、讽刺标记词。  
语义–情绪不一致度：情感分与语义极性冲突的量化指标。  
  
3.4 建模设计  
三分类模型：文明（0）/伪文明（1）/不文明（2）。  
两阶段模型（备选）：文明vs非文明；再识别伪文明。  
  
3.5 统计分析与可视化  
聚类可视化、组内组间相似度、差异检验（Mann-Whitney U）、SHAP解释。  
  
3.6 评价指标与成功标准  
主指标：伪文明类别的F1分数提升≥+5pp。  
次指标：整体Macro-F1、文明误杀率与不文明漏检率。

## 四、预期发现与局限（Findings & Limitations）

4.1 预期发现  
- 加入“礼貌强度+情绪不一致度”后，伪文明识别显著改善。  
- 伪文明在语义空间靠近文明，在情绪维度偏向不文明。  
- 典型伪文明特征组合：正向词+负向情绪/讽刺标记+反问+高不一致度。  
  
4.2 可能局限  
- 主观性：伪文明判定存在文化差异。  
- 域迁移：跨平台伪文明模式异构。  
- 上下文缺失：缺乏对话语境削弱讽刺识别。  
- 对抗适应：攻击者可能进化出更隐蔽的伪装。