# 伪文明评论研究通俗版说明

## 一、研究背景：有些“不文明”穿着“文明外衣”

在网络评论里，不文明的内容并不总是脏话或直接辱骂。  
很多人会用一种“假礼貌”的方式表达讽刺或攻击，比如：  
  
“谢谢你又一次证明你什么都不懂。”  
“真聪明啊，能想出这种主意。”  
  
这些评论看起来有“谢谢”“真聪明”这样的好听词，但语气其实是在嘲讽对方。  
我们把这种现象叫做“**伪文明评论”**——表面文明、语气克制，但其实带着情绪或攻击意图。  
  
在我们的数据里，很多这样的评论被误判为“文明”，  
因为机器模型只看字面意思，看不懂“语气里的讽刺”。  
所以我们希望找出这类评论的规律，让模型能更聪明地识别它们。

## 二、研究思路：礼貌 + 情绪 = 揭开“假文明”面具

我们提出的思路很简单：不仅看“评论里说了什么”，还要看“评论的语气和情绪是什么样的”。  
  
具体来说，我们从数据里提取了两类指标：  
  
1. 礼貌强度（Politeness Score）  
统计评论中礼貌词（如 please, thank, appreciate）的数量，看它“外表有多文明”。  
  
2. 情绪不一致度（Emotion–Semantic Inconsistency）  
比较一句话的字面意思和真实情绪是否一致。  
如果表面积极但情绪负向（比如带嘲讽的“谢谢”），不一致度就高。  
  
我们发现，这两个指标的组合特别能区分出“伪文明评论”。

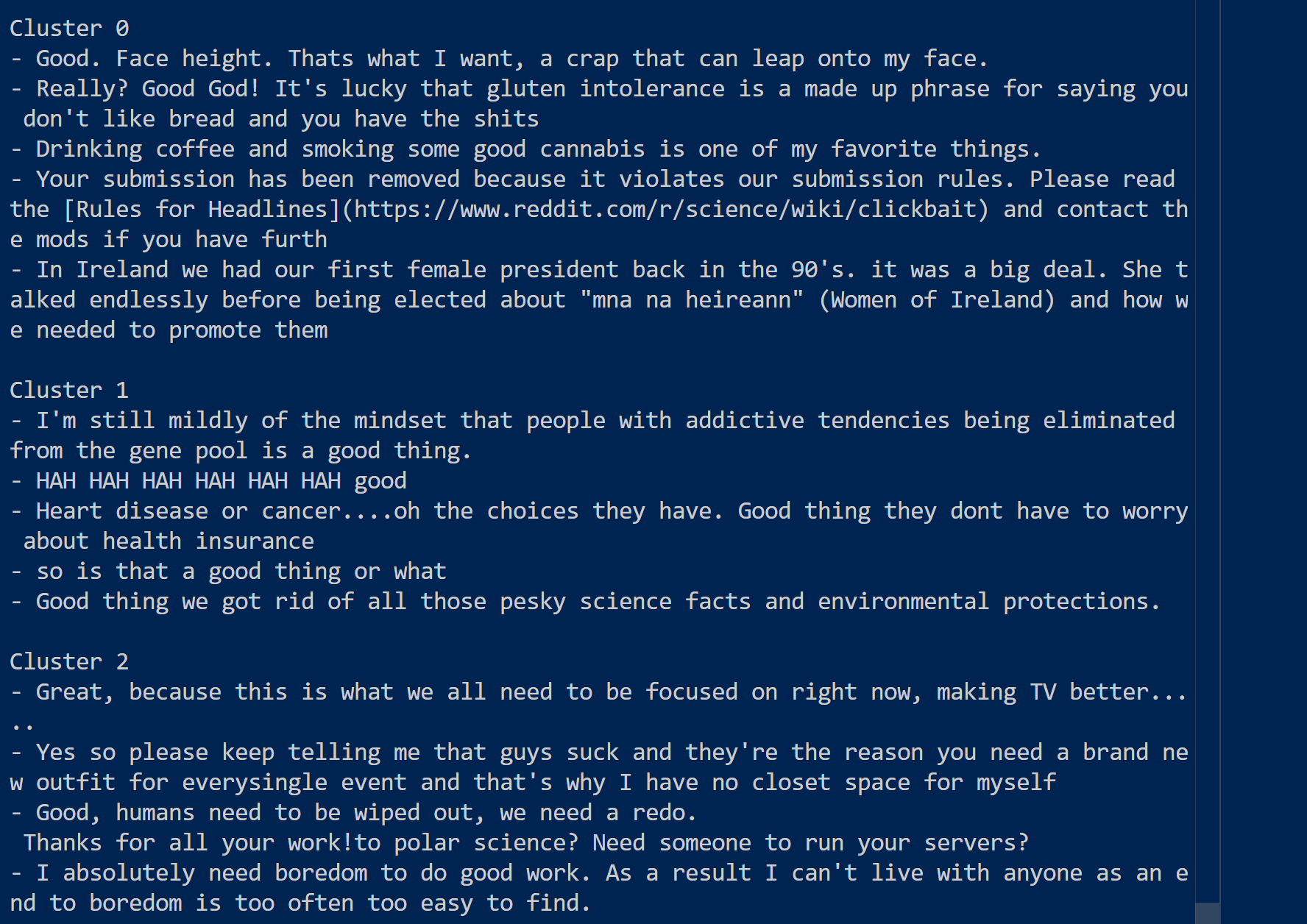
## 三、数据和模型：怎么做的

我们使用了一个含有上万条评论的数据集，每条评论有是否被移除的标签：  
  
类别 | 示例 | 特征  
---- | ---- | ----  
文明 | “Thank you for your explanation.” | 礼貌高、情绪正向  
伪文明 | “Thank you for ruining everything again.” | 礼貌高、情绪负向  
不文明 | “You are so stupid.” | 礼貌低、情绪负向  
  
我们用句子向量模型（Sentence-BERT）提取语义，再加上上面的两个新特征。  
训练后模型能自动预测评论属于哪一类。  
**根据假设进行数据分析**

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 + 语义–情绪不一致度”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）

你们的任务是

1. 我之前的代码是用的小数据，大概1万多条，看能不能把数据量扩大
2. 我已经用聚类发现了一些例子，共有 4 个簇（sizes: 642, 27, 16, 19）——第 0 簇占多数（642/704），top\_terms 显示很多数字/年份/链接样式的token（表明许多文本包含帖子格式、时间、编号或引用），其它簇包含短文本、情绪化或带“good/thank”但语境可能负面或讽刺的表达。你们需要找到语义向量和情绪指标的关系，证明H1（H2是H1的延申）H3可选
3. 最后是可视化结果



## 四、主要发现

我们的实验结果显示：  
  
模型版本 | 特征 | 伪文明识别能力（F1分数）  
---- | ---- | ----  
原始模型 | 只看语义 | 识别一般（容易混淆文明和伪文明）  
改进模型 | 加入“礼貌强度 + 情绪不一致度” | 明显提升，能抓住假礼貌攻击  
  
简单来说，加上这两个特征后：  
- 模型能更准确地区分“真文明”和“假礼貌”；  
- 不会再被“谢谢”“真聪明”这种讽刺性语句迷惑；  
- 文明评论的误删率下降了，伪文明评论的识别率上升了。

## 五、这说明了什么？

语言不只是字面意思，还藏着情绪。  
我们的模型如果只看“说了什么”，会被伪装骗到。  
要让它更聪明，就得同时理解“怎么说”的部分。  
  
伪文明是情绪和理性的混合体。  
它既带着情绪（不满、讽刺），又用理性的表达包起来。  
这让它在语义上更接近文明，但在情绪上偏向不文明。  
  
平台治理需要灰色地带的识别。  
不文明 ≠ 脏话。我们找到一种新型风险语言：看起来文明，但暗藏攻击。

## 六、局限与下一步计划

人工判断仍有主观性：讽刺和幽默有时难区分，不同人理解不一样。  
主题差异：政治话题下伪文明更多；娱乐话题相对少。  
模型仍需改进：缺乏上下文时，有些语句的语气难判断。  
  
下一步我们会：  
- 让模型学会“语境理解”，判断语气是认真还是讽刺；  
- 测试不同社交平台上的语言差异；  
- 优化模型解释功能，让内容审核更公平、更透明。

## ✅ 总结一句话

伪文明评论看起来礼貌，其实暗含攻击。  
我们用“礼貌强度 + 情绪不一致”两个指标，  
让模型学会分辨“真礼貌”和“假礼貌”，  
让机器更懂人话，也让网络更友善。