跨文化冒犯性评价的受众差异与统一内容审核模型的适用边界——基于 D3CODE 数据集的可解释机器学习研究

陈祎来

（宁波大学阳明学院，浙江宁波，315211）

摘要：全球平台常以统一模型处理不同文化背景受众的内容评价，但同一表达在不同地区和个体之间可能获得截然不同的冒犯性评分。现有研究多关注文本特征，而对受众地区与个体价值的影响，以及模型在未见地区中的表现均讨论不足。本研究基于 D3CODE[1] 数据集的 4554 条英文评论、 4309 名标注者和 144730 条有效评分，通过消融比较、留一地区外验证、误差切片与 SHAP 分析，考察文本、地区和道德价值对冒犯性评价预测的影响。研究发现，受众地区带来的预测改善高于个体价值；同时纳入地区、价值与预设交互关系的模型表现最佳，但个体价值并未降低八个地区的宏平均误差。评论分歧越大，模型预测误差越高，宗教、社会群体、性别及关系规范等内容也更易出现偏差。研究表明，冒犯性评价并非仅由文本决定，统一内容审核模型难以充分覆盖跨文化受众差异。平台除报告总体性能外，还应分别检验不同地区、议题与分歧水平下的模型表现。

关键词：跨文化传播；冒犯性评价；受众差异；平台治理；内容审核；计算传播

中图分类号：G206　　**文献标识码：**A

一、问题的提出

(一)冒犯性评价随情境而变

全球社交媒体每天面对数量庞大、语境交叠的用户表达。为维持规模化运行，平台通常把冒犯性转化为可识别的规则、标签和分数，并据此安排内容筛查、队列排序与处置。这样的技术处理提高了审核效率，也容易把冒犯性固定为文本自身的属性。事实上，受众对同一表达的判断会随互动关系、规范期待和所处情境而变化；脱离具体传播过程，词语和句式只能提供线索，难以穷尽冒犯评价的形成条件。

从语言使用来看，不礼貌研究早已注意到面子需求、社会规范、情绪和互动语境对冒犯评价的共同作用，同一种表达在不同关系和场合中可能获得不同理解[2]。这一认识把研究重点从识别若干冒犯词语推进到考察评价如何在具体交往中形成。D3CODE 让来自不同地区的受众平行评价同一批评论，文本由此成为共同刺激，评分差异则呈现受众判断的变化。相同文本之下仍然存在分散评价，说明跨文化冒犯判断需要同时考察内容与受众。

从传播过程来看，霍尔的编码 / 解码模式指出，受众会调动知识框架、社会位置和文化资源理解媒介内容[3]；Morley 的受众研究进一步表明，社会位置会塑造可用的解释资源，却不会把个人理解机械地归入同一种模式[4]。地区位置与个体价值因而构成两个不同层次的受众表征：前者概括共同环境，后者呈现个人规范取向。基于 D3CODE 的数据形态，本研究把分析重心放在评分层面的受众差异，生产者意图与自然接收过程则留待情境性材料进一步考察。由此，问题不再停留于某条评论是否冒犯，而进一步转向：在文本保持相同的情况下，

哪些受众信息能够表征并预测评分差异？

(二)全球平台需要聚合多元判断

当不同背景的受众汇入同一平台，冒犯判断的情境差异便会转化为全球内容治理的现实难题。Marwick 和 boyd 将多重受众同时进入共同传播场景的现象概括为语境坍塌，内容生产者由此需要面向边界模糊的想象受众表达[5]。Davis 和 Jurgenson 进一步区分了主动促成的语境合流与非预期发生的语境碰撞[6]。在全球平台上，语言习惯、宗教背景、身份经验与规范期待彼此交叠，同一内容更容易跨越原有语境，也更可能获得分化的评价。

然而，监督式审核模型通常依赖经过编码或聚合的训练标签。平均值、多数表决或统一阈值降低了操作复杂度，也重新分配了不同判断的可见程度：接近中心趋势的意见更容易被保留，持续存在的少数评价则可能受到压缩。为把这一聚合逻辑转化为可观察对象，本研究以单值文本分数作为分析基线，考察共同预测与个人评分之间的偏离。该基线服务于误差分布分析，使统一输出在哪些地区、议题和分歧情境中更容易失准得以呈现。

内容审核的意义也因此超出分类性能。Gillespie 把内容审核视为平台组织公共话语、设定表达边界的核心活动[7]；当自动化系统进入这一过程，标签如何形成、规则如何编码、误差如何分布，都会转化为平台治理问题[8]。国内智能传播研究同样指出，数据选择、规则设定和运算过程可能共同生成算法偏向[9-10]。由此，对统一审核模型的评估既要考察总体预测表现，也要追踪误差如何分布于不同受众和内容情境。把被平均值遮蔽的差异重新呈现出来，才能进一步识别需要补充样本、调整规则或引入人工复核的具体位置。

(三)地区与价值提供不同受众信息

围绕多元判断如何进入模型，既有研究首先转向标注者差异。语义任务中的分歧往往包含合理的解释差异；若直接采用多数表决，少数视角和模型不确定性会随标签聚合一同消失[11-12]。进一步的经验研究发现，标注者身份、既有信念和人口背景与主观评分相关，不过这些因素的作用会随内容和任务而变化[13-14]。主观评价任务中的标注记录由此不宜被先验视为完全可互换的数据，而应成为理解评分差异的重要对象。

在现有研究中，地区与个体价值构成两类常见的受众指标。地区标签能够简洁概括共同生活环境，却也把国家、语言、制度、宗教和平台经验等多重因素压缩在同一类别中；道德基础问卷则从个人层面测量关怀、平等、比例、权威、忠诚和纯洁等价值关切[15]。D3CODE 原研究发现，冒犯评分在地区之间存在差异，部分评分差异同道德价值相关，价值聚类与地理区域又未完全重合[1,16]。另有研究尝试以国家文化特征预测冒犯语言任务的跨域迁移效果[17]。这些发现共同表明，地区主要呈现群体环境，价值问卷深入个人规范取向，两类指标分别保留了受众差异的一个侧面。

不过，既有研究较多描述地区、价值与冒犯评分的样本内关联，对于二者在共同基线上的增量预测效度、面向未见地区的迁移表现，以及误差在不同受众和内容切片中的分布，仍

缺少系统比较。若同一评论或同一标注者跨越训练集和测试集，评估便会混入重复观察，弱化对新评论和新受众的外推检验[18]。在总体误差之外引入地区、议题和评分分歧切片，则可以进一步呈现模型误差的分布结构。由此，地区标签与个体价值究竟提供多少增量信息、这些信息能否跨地区迁移，以及单值文本预测的误差集中在哪里，构成三个彼此关联而又需要分别检验的问题。

计算传播研究可以把上述问题转化为可检验的经验命题。王成军把可计算化视为传播研究连接大规模数据、模式与机制的重要方式[19]；李雪莲、刘德寰及 Hofman 等进一步区分了预测与解释两类认识任务[20-21]。沿着这一思路，本研究通过共同基线上的特征消融比较地区与价值的增量预测信息，以留一地区外验证考察外部适用范围，并借助分组误差和 SHAP 追踪模型实际依赖的特征。这样的设计把既有描述性差异推进到样本外预测与跨地区检验，也为后续讨论统一审核模型的适用边界提供经验依据。

二、研究设计

围绕地区位置与个体价值的预测增益及跨地区效度，本研究依次完成数据整理、模型比较和解释分析。首先，将 D3CODE 整理为评论、标注者和个人评分三个相互关联的层次，并据此构造文本、内容、人口与价值特征。其次，在共同基线上比较地区和价值带来的样本外预测信息，再将目标地区整体移出训练集，考察个体价值的跨地区表现。最后，通过分组误差与 SHAP 分析呈现单值文本预测的误差分布以及联合模型的特征依赖。

(一)分析框架与研究问题

本研究从文本、个体和地区三个层次组织变量。评论提供共同的语言形式和议题线索；年龄、性别、社会经济地位与道德价值记录标注者的个人信息；地区位置概括群体生活环境及相关经验。结果变量为标注者对评论给出的 0—4 级冒犯性评分。

地区与价值对应两种层次不同的受众表征。地区标签概括语言环境、制度经验、宗教结构和平台接触等共同条件，同时也会压缩地区内部差别；道德价值问卷在个人层面记录连续分数，测量更为细致，其经验含义则受到问卷维度、使用语言和具体议题的制约。因此，两类变量在本研究中承担文化相关受众信息的观察窗口，而非完整的文化测量。

经验分析分为两个层次。已有地区中的消融比较在同时留出新评论和新标注者的条件下，考察地区与价值相对于共同基线增加了多少预测信息；跨地区验证则把目标地区整体移出训练集，检验价值信息能否迁移到新的群体环境。另以单值文本模型作为分析基线，观察统一分数与个人评分之间的误差如何分布。由此，模型比较集中考察两类信息的预测增益及其外部适用范围，并为后续机制研究标出值得追问的受众差异。

据此，本研究提出以下三个研究问题：

RQ1：在共同基线上，地区位置与个体道德价值的增量预测效度有何差异？

RQ2：当目标地区完全未参与模型训练时，个体道德价值能否改善样本外预测表现？

RQ3：单值文本基线的预测误差如何分布于不同地区、内容类型和评分分歧水平？

(二)数据来源与变量构造

1. 数据来源与分析单位

研究数据来自 D3CODE，其中的评论从 Jigsaw 语料中按随机抽样、道德表达抽样和社会群体相关内容抽样三种策略选取[1]。原始评论表共 4590 行，部分评论因同时属于多个类别而重复出现。本研究按评论编号合并重复记录并保留多类别标识，得到 4554 条唯一评论。标注者表包含 4309 人，覆盖 21 个国家和 8 个地区；评分表共有 150702 条“评论—标注者”记录。

原始量表采用 0—4 五级编码，其中 0 表示完全不冒犯，4 表示极其冒犯，-1 表示标注者未理解评论。剔除 5972 条未理解记录后，主分析保留 144730 条有效评分。每条评论接受多名标注者评价，每名标注者又评价 34 或 35 条评论，由此形成评论与标注者交叉重复的观察结构。后续结构化模型围绕文本、内容、人口、地区与价值五类变量展开。标注者编号只承担个体识别功能，纳入训练会使模型记忆个人；国家字段与地区高度重叠，进一步细分后样本也更稀疏，因而均不作为预测特征。统一定义实验字段在公开数据中无法稳定对应，亦未进入分析。

2. 文本基础分与内容特征

文本特征分两阶段处理。第一阶段在评论层训练词频—逆文档频率（TF-IDF）与岭回归模型，词项由一元词和二元词组构成，最多保留 20000 个特征。训练目标先计算每条评论在各地区的平均评分，再对地区均值等权平均，以减轻不同地区标注数量的影响。每条训练评论通过嵌套折外预测获得文本基础分；外层测试评论则由未见过该评论的模型评分。经过这一处理，高维词项被压缩为一个可进入结构化模型的文本分数，同时维持评论层面的样本外预测。

第二阶段采用少量可解释的内容变量，包括文本长度、感叹号数量，随机抽样、道德表达和社会群体相关三类上层标签，以及宗教、性别与关系规范两类细分标签。评论可以同时具有多个内容标签，后续内容切片按照这种多重归属进行描述性比较。

3. 标注者与价值特征

标注者特征包括年龄、性别、社会经济地位和地区。个体价值来自关怀、平等、比例、权威、忠诚和纯洁六项道德基础。为同时区分整体认同水平与价值结构，本研究先计算六项价值的个人均值，再以各维度减去这一均值，得到相对重视程度。六个中心化值之和恒为零，模型因此保留关怀、平等、比例、权威和纯洁五项，以忠诚作为参照。最终进入模型的是道德价值总体均值与五项相对值，从而避免原始值、均值和中心化值之间的重复表达。

内容与价值的交互依据理论对应预先限定为三项：纯洁相对值与宗教相关内容、平等相

对值与性别及关系规范相关内容、关怀相对值与社会群体相关内容。这样的设置使交互分析集中于具有明确含义的关系，并保持模型规模适合后续解释。

(三)模型设定与检验策略

1. 消融模型设置

消融实验由 T、B、R、V 和 RV 五组模型构成。T 模型直接采用岭回归生成的文本基础分，代表同一评论对应一个共同分数的文本基线。B 模型在文本分数上加入内容形式、抽样类别、年龄、性别和社会经济地位；R 与 V 分别在 B 的基础上加入地区和个体价值；RV 则同时加入地区、价值与三项预设交互。B、R、V、RV 统一使用以绝对误差为目标的梯度提升树[22]，并固定学习器与参数，使模型之间的差异对应新增特征组。结构化模型的预测在评价前截断至 0—4 的量表范围。

T 学习评论层、地区等权的聚合目标，B、R、V 和 RV 则预测“评论—标注者”层面的个人评分。由 T 到 B 还同时改变了学习器、损失函数和特征规模，因此 T 用于提供技术基准并开展误差切片；对地区与价值增量效度的受控比较，则从共同基线 B 出发，在 R、V 与 RV 之间展开。

2. 双重留出与评价指标

主实验同时留出评论和标注者。研究先将评论随机分为五折，再在每个地区内部将标注者分为五折。第 k 折训练集使用评论折与标注者折均不等于 k 的记录，测试集则由二者均等于 k 的记录组成。测试范围由评论折与标注者折相等的五组对角记录构成，合计 29161 条，每一折均覆盖八个地区。模型据此面对训练阶段未出现的评论和标注者，测试结果反映两类新对象同时出现时的预测表现。

主模型评价以平均绝对误差（MAE）为中心，同时报告最差地区 MAE 以及评分与预测的 Spearman 等级相关；绝对误差不低于 2 分的严重误差率用于后续分组诊断。表 1 中的总体 MAE 和 Spearman 相关均取五个测试折的均值，最差地区 MAE 则先在每折取八地区中的最大值，再跨折平均。模型差异的区间估计先在 4521 条具有配对预测的评论内，平均多名标注者的绝对误差，再以评论为单位进行 2000 次配对自助抽样。前者以测试折为汇总层次，后者赋予每条评论相同权重，两种统计量分别呈现整体表现与配对差异；自助抽样区间对应既定数据划分和模型训练下的评论层配对差异。

3. 跨地区验证

跨地区实验每次将一个地区完全移出训练集，并同步移除地区字段，比较共同基线 B 与价值模型 V。B 和 V 均不把地区作为直接输入；文本基础分的训练目标仍在其余训练地区内先求地区均值，再作等权汇总。每个目标地区再按评论执行五折外测，使测试评论在对应训练中同样保持未见。八个地区的预测合并后，B 和 V 均覆盖 144730 条有效评分一次。研究据此比较各地区 MAE、八地区宏平均 MAE、最差地区 MAE 与平均 Spearman 相关，观

察个体价值能否在新的群体环境中稳定降低误差。

4. 模型解释

对于主实验中 MAE 最低的结构化模型，本研究进一步开展 SHAP 分析。具体做法是分别解释五个外层折中拟合模型的原始输出，再合并各折 SHAP 值；评价指标使用的则是截断至 0—4 后的预测。SHAP 将原始输出相对模型基准值的变化分配到各项输入特征，可据此观察模型主要依赖哪些信息以及贡献方向如何变化[23]。为平衡地区构成，本研究在每个外层折、每个地区以固定随机种子抽取至多 75 条严格外测记录，五折合计形成 3000 条解释样本，每个地区各有 375 条。总体蜂群图与分地区热力图均据此生成。

局部瀑布图采用目的性案例。候选评论需要在八个地区各有至少 3 条评分，总体评分覆盖 0—4，评分标准差不低于 1；在此基础上，再从严格测试记录中选择真实评分相差至少 3 分，且价值特征和模型预测均有差异的跨地区标注者。总体图和地区图呈现模型在解释样本中的整体依赖与地区差异，局部案例则展示相同文本下各项特征如何共同形成两次预测。SHAP 在本研究中解释模型计算过程，相关数值统一表述为特征对预测的贡献。

三、研究发现

(一)地区信息在已有地区中的预测增益更高

首先，从整体表现看，单值文本基线 T 的 MAE 为 1.142，Spearman 相关系数为 0.193；加入内容与基本人口特征后，B 模型的 MAE 降至 1.026，相关系数升至 0.243（见表 1）。由于两组模型同时改变了预测层级、学习器、损失函数和特征规模，这一变化呈现的是从评论共同分数到个人结构化预测的整体差别。地区与价值的增量比较以 B 为共同起点。

表 1 不同特征组合的双重留出测试结果

模型	总体构成	总体 MAE	最差地区 MAE	Spearman
T	文本基础分	1.142	1.269	0.193
B	T＋内容与人口共同基线	1.026	1.207	0.243
R	B＋地区	1.018	1.190	0.268
V	B＋个体道德价值	1.025	1.190	0.265
RV	B＋地区＋价值＋预设交互	**1.013**	**1.175**	**0.279**

进一步比较三组受众模型，R 的折均 MAE 为 1.018，低于 V 的 1.025。按评论等权计算，R 相对 B 的配对平均绝对误差差值为 -0.0105，95% 置信区间为 [-0.0143,-0.0067]；V 相对 B 为 -0.0040，95% 置信区间为 [-0.0074,-0.0005]。两类受众信息都改善了共同基线，但地区带来的降幅更大。V 相对 R 的配对差值为 0.0066，95% 置信区间为 [0.0029,0.0103]，也说明在已有地区的新评论和新受众中，地区标签包含了更多样本外预测信息。

RV 模型取得五组模型中的最佳表现，其折均 MAE 为 1.013，最差地区 MAE 为 1.175，Spearman 相关系数为 0.279。按评论等权计算，RV 相对 B 的配对平均绝对误差差值为 -0.0178，95% 置信区间为 [-0.0223,-0.0135]；相对 R 为 -0.0073，95% 置信区间为 [-0.0105,-0.0042]。地区、价值和内容—价值交互共同进入后，模型因而获得了额外的预测信息。

这一结果需要结合变量层次理解。地区标签可能同时汇集语言环境、制度经验、宗教结构、英语使用方式和样本中相对稳定的评分差别，其较高效度反映了综合群体信息，而非区域内部文化同质性的证据。与此同时，RV 相对 R 同步增加了价值主效应和三项预设交互，其增益对应整个联合特征组；价值主效应与交互项各自贡献多少，还需要更细的组内消融。就 RQ1 而言，当前证据表明地区信息在已有地区中的增益高于个体价值，而联合使用两类受众表征获得了最低预测误差。

(二)联合模型主要依赖文本与内容特征

RV 模型在主消融中表现最好，本研究据此对地区均衡抽取的 3000 条严格外测记录开展 SHAP 分析。总体上，模型首先依赖文本与内容信息。图 1 按平均绝对 SHAP 值排列前十五项特征，文本基础分居首(0.118)，道德表达内容次之(0.107)；道德价值总体均值(0.072)、文本长度（0.053）和社会经济地位（0.040）随后进入前五。社会群体相关内容及其与关怀相对值的交互也位于前列。地区变量中，大洋洲、阿拉伯文化区和汉字文化圈的贡献较为突出。

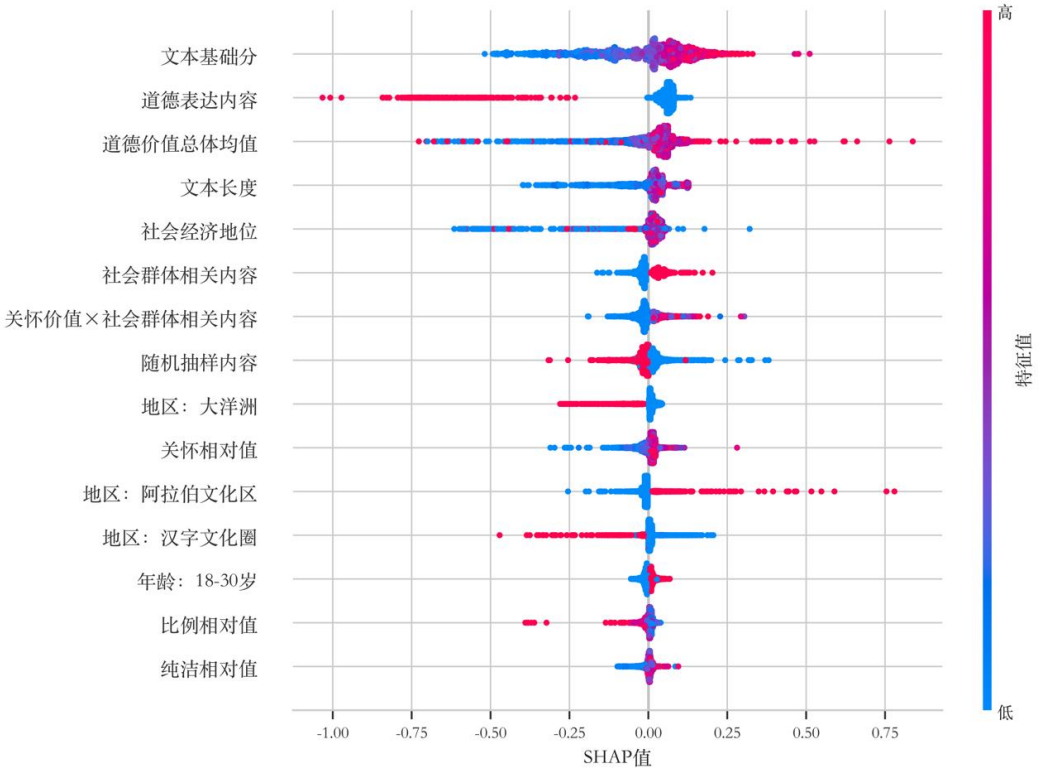

图 1　地区均衡外测样本中的模型特征依赖

蜂群图同时呈现了特征取值与贡献方向。文本基础分升高通常推动预测分数上升；道德表达标签取 1 的样本，其 SHAP 值整体偏负，社会群体相关标签取 1 的样本则整体偏正。道德价值总体均值和地区变量在横轴两侧均有分布，说明其贡献方向随样本组合而变化。联合模型通过树模型中的条件关系组织多类输入，价值与地区在不同样本中因而具有不同的贡献方向。

总体排序之外，模型依赖的特征构成在地区之间也有差别。图 2 将同一解释样本按地区计算平均绝对 SHAP 值。文本基础分和道德表达内容在多数地区都处于前列，但局部排序并不完全相同。阿拉伯文化区中，道德价值总体均值和本地区标签的平均贡献分别为 0.093 和 0.075；大洋洲中，文本基础分达到 0.134，本地区标签为 0.080，关怀价值与社会群体相关内容的交互为 0.037；汉字文化圈中，社会经济地位的平均贡献为 0.109，高于文本基础分的 0.098。地区之间的差别由此表现为模型实际调用的信息组合不同，而不只是总体误差高低不同。

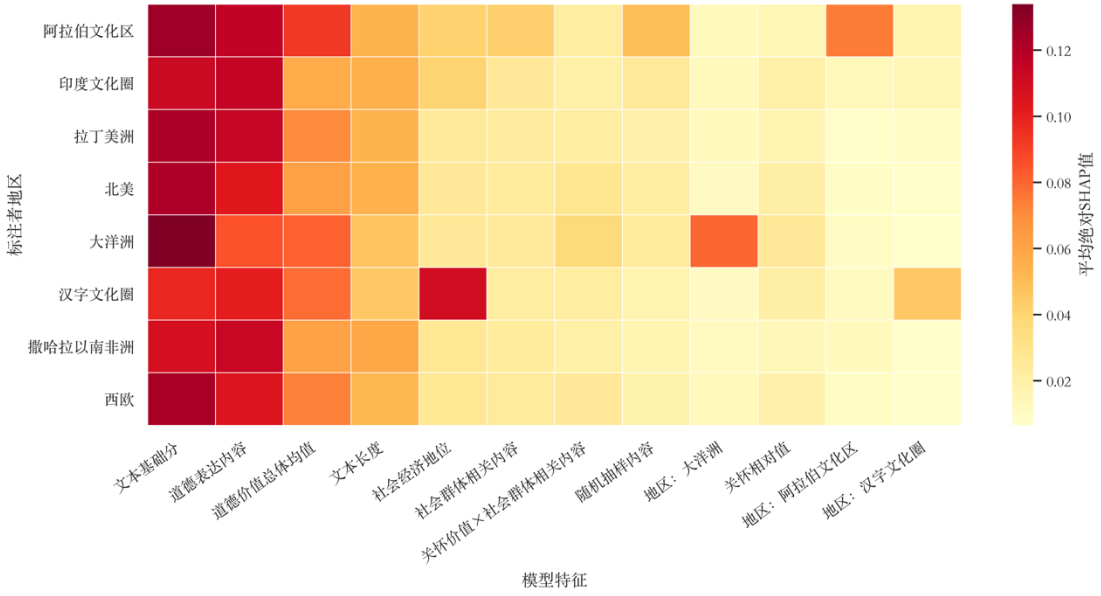

图 2　不同地区的平均绝对 SHAP 值

最后，局部解释展示同一文本如何得到不同预测。图 3 选择了一条涉及宗教与政治讽刺的评论。阿拉伯文化区标注者 A 给出 4 分，模型预测 3.62；大洋洲标注者 B 给出 1 分，模型预测 1.00。两人的文本基础分完全相同，预测差距来自这一共同分数与不同受众特征在非线性模型中的组合。对 A 而言，道德价值总体均值、阿拉伯文化区标签和年龄特征分别带来较明显的正向贡献；对 B 而言，大洋洲标签和较低的纯洁相对值略微压低预测。这组目的性案例补充了前两幅总体图，显示模型如何在具体记录中组合特征。

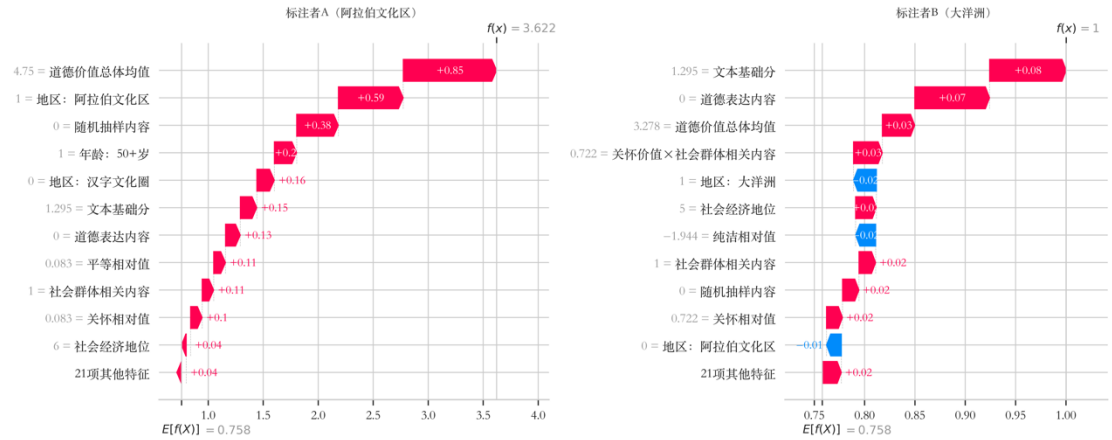

图 3　同一评论、不同标注者的局部解释

三组 SHAP 结果分别呈现总体依赖、地区构成和局部计算过程。由此可见，地区与价值并非以固定方向进入预测，而是在具体文本和受众组合中共同发挥作用。这里的贡献属于模型计算层面的依赖；若要进一步说明受众为何形成相关判断，还需结合开放式理由、访谈或实验材料。相关特征间的贡献也可能相互分摊，因此单项排序更适合识别模型依赖，不能直接等同于心理机制。

(三)个体价值未形成稳定的跨地区优势

已有地区中的预测增益能否延伸至新的地区，需要进一步接受外部检验。留一地区外实验将目标地区整体移出训练集，同时移除地区字段，在共同基线 B 与价值模型 V 之间展开比较。这样既避免用目标地区名称参与预测，也使价值信息直接面对训练阶段未出现的群体环境。八个地区的外测结果见图 4。

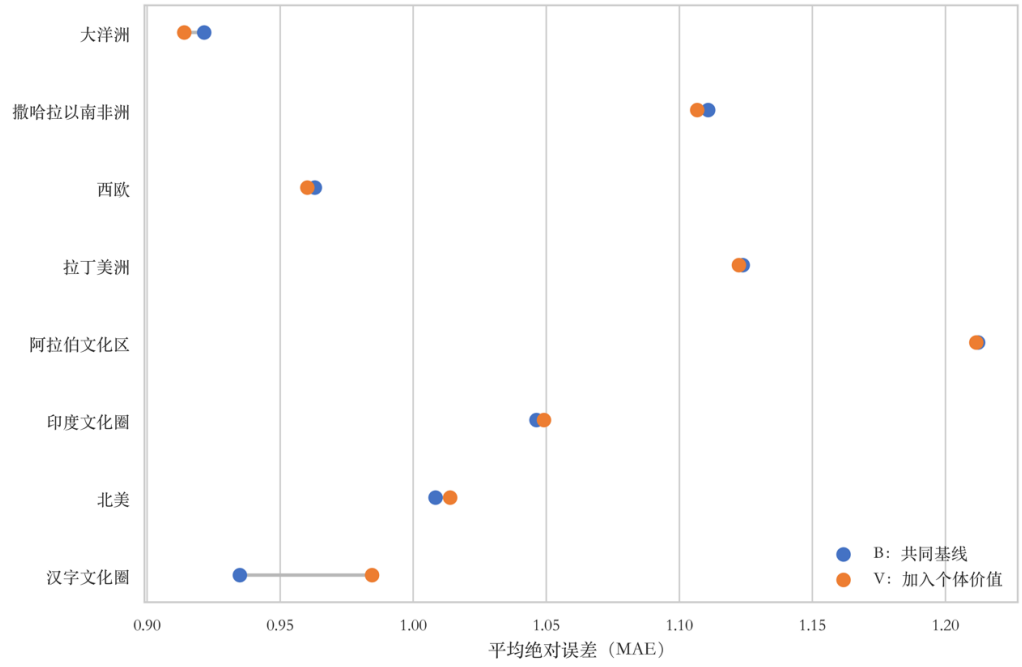

图 4　个体价值在完全未见地区与评论上的预测表现

从宏平均结果看，B 的 MAE 为 1.040，V 为 1.045，加入价值后整体误差略有上升。最差地区 MAE 从 1.2122 降至 1.2116，变化幅度很小；平均 Spearman 相关系数则由 0.209 降为 0.204。个体价值因而没有在三个总体指标上形成一致改善。

分地区结果以描述性比较呈现出更细的差别。V 在阿拉伯文化区、拉丁美洲、大洋洲、撒哈拉以南非洲和西欧降低 MAE，在印度文化圈、北美和汉字文化圈提高 MAE。五个改善地区的降幅均较小，其中大洋洲最大，为 0.0075；汉字文化圈的 MAE 则上升约 0.050。个体价值在部分目标地区提供了局部信息，这些收益没有汇聚为八地区的稳定优势。

五个地区的小幅改善与三个地区的误差上升并存，说明价值量表的跨地区效度取决于目标地区的问卷含义、内容构成、英语使用经验和评分关系，而非仅由变量处在个人层面所决定。留一地区外实验把这些差异完整保留在目标环境中，因而更接近模型迁移时面对的条件。就 RQ2 而言，个体价值尚未表现出稳定的跨地区优势；文化敏感建模若引入此类字段，仍需先检验量表与评分关系在目标地区是否可比。

(四)文本基线误差在地区、议题与分歧层面分化

RQ3 转向单值文本预测的误差边界。本研究将 T 模型的 29161 条严格测试记录按地区、内容和评论分歧程度展开，评论分歧依据全部有效评分的标准差划分四分位数。该变量在模型训练完成后用于误差诊断，图 5 中的虚线表示 T 模型总体 MAE 1.142。

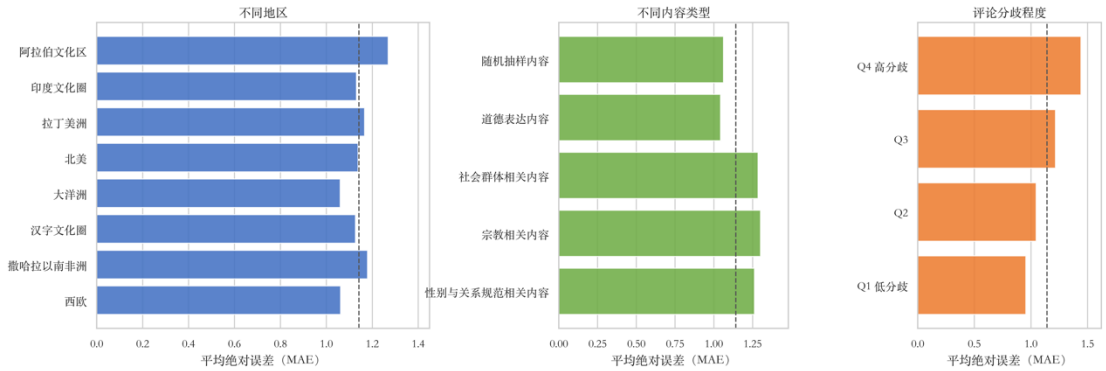

图 5　单值文本基线的描述性误差切片

一方面，地区之间的误差存在明显差别。阿拉伯文化区 MAE 最高，为 1.269；撒哈拉以南非洲和拉丁美洲分别为 1.180 和 1.166，大洋洲与西欧均约为 1.06。这组数值呈现 T 模型与个人评分之间的实际偏离，其中同时包含地区评分分布和样本构成的差别，适合用于判断哪些地区需要优先接受进一步审计。

另一方面，内容切片同样呈现清晰分化。宗教相关内容 MAE 为 1.301，社会群体相关内容为 1.285，性别与关系规范相关内容为 1.262，均高于总体水平；随机抽样和道德表达内容则约为 1.06 和 1.04。一条评论可以同时归入多个标签，因此，这些数值描述的是各类内容切片的风险集中程度，而不是互斥类别之间的净差异。较高误差使宗教、社会群体与性别及关系规范相关内容成为优先审计对象；组间不确定性及样本构成的影响，则可在后续

推断分析中继续检验。

此外，模型误差会随评分分歧扩大。低分歧 Q1 的 MAE 为 0.955，Q2 为 1.046，Q3 升至 1.217，高分歧 Q4 达到 1.441；严重误差率也从 Q1 的 1.7% 增至 Q4 的 24.5%。这一递增关系与任务结构密切相关：T 模型为同一评论生成一个共同分数，而分歧四分位描述个人评分的离散程度。评分越分散，单一输出便越难同时贴近不同受众。分歧梯度由此揭示单值预测覆盖离散评价的能力，其文化和语义来源则需要更多情境材料加以辨认。

综合来看，T 模型的总体误差掩盖了三个层面的差异：部分地区与议题更容易出现较大偏离，高分歧评论则集中承载严重误差。地区和内容切片指向优先审计对象，分歧梯度呈现单值输出的结构性边界。评分分歧还可能包含语言理解、个人经验和标注质量等信息，后续研究可据此选择样本，进一步考察不同受众如何形成彼此分化的冒犯判断。

四、结论与讨论

本研究以“评论—标注者”配对记录为分析单位，在同时留出新评论和新标注者的条件下，比较地区位置与个体道德价值对冒犯性评分的增量预测信息，并以留一地区外验证考察两类受众表征的迁移表现。研究发现集中在三个相互关联的方面：文本特征构成预测的主要依据，受众信息能够进一步降低个体评分误差；地区标签在既有地区中的预测增益高于个体价值，地区、价值与预设交互共同进入后，RV 规格取得最低误差；当目标地区完全退出训练集时，个体价值虽在五个地区带来小幅改善，却未降低八地区宏平均误差。与此同时，单值文本基线的误差在不同地区和内容议题之间存在差异，并随评分离散度升高，由此呈现出该基线在当前任务中的误差边界。

首先，跨文化冒犯性评价同时包含文本倾向与受众差异。在 RV 模型的解释样本中，文本基础分的平均绝对 SHAP 值居首，评论措辞由此构成模型预测的重要线索；R 与 V 又分别在共同基线 B 上降低误差，说明受众位置与价值取向也包含额外预测信息。Culpeper 对不礼貌的讨论强调社会规范和具体语境，霍尔的编码 / 解码框架则指出，传播文本的意义需要在受众解释中完成[2-3]。上述预测结果与这一认识相呼应：文本提供共同刺激，受众位置与价值取向补充评价分化的信息。由此来看，标注分歧保存着社会位置、规范期待和传播情境参与判断的痕迹，应作为需要解释的受众变异进入分析。聚合多人评分时，研究者据此需要交代标注者构成与聚合规则，并检视平均值是否遮蔽了持续存在的少数意见。

其次，地区标签与个体价值分别呈现群体环境和个人取向，其预测作用处在不同测量层次，也受各自测量边界制约。在域内严格外测中，R 规格的 MAE 为 1.018，优于 V 规格的 1.025；RV 规格进一步降至 1.013，表明联合规格包含额外预测信息。地区标签可能汇集语言环境、宗教结构、制度经验、英语使用方式和平台接触等多重因素，信息范围较宽，也更贴近当前样本的群体结构；六维道德价值则深入个体层面，有助于辨认地区内部的差别。由

于 RV 相较 R 同步加入价值主效应和三项预设交互，现有消融可以确认联合规格的改善，却尚难把增益进一步分解到某一组新增特征。留一地区外实验则显示，加入价值后宏平均 MAE 从 1.040 升至 1.045，五个地区的局部改善未能转化为稳定的整体收益。地区标签在既有地区中表现较好，更适合被理解为多种环境信息的综合表征；个体价值提高了域内分辨率，却尚未形成可重复的跨地区迁移优势。因此，文化表征是否有效，需要同时接受域内增益和域外迁移两种检验。

再次，单值文本基线的误差分布揭示了平均性能之外的审核难题。T 模型在阿拉伯文化区、撒哈拉以南非洲和拉丁美洲的误差相对较高，在宗教、社会群体以及性别与关系规范相关内容上的 MAE 也高于总体水平。更为明显的是，评分分歧从低分位升至高分位时，MAE 由 0.955 增至 1.441，严重误差率则由 1.7% 增至 24.5%。评分越分散，单一预测值越难同时贴近不同受众，这一结构特征揭示了单值输出在意义分化情境中的基本困难，也为理解统一审核的潜在风险提供了线索。作为描述性诊断，地区、议题和分歧切片适合用来定位优先审计对象，具体成因则有待结合样本构成与受众解释继续分析。平台评估若只报告总体 MAE，容易把局部误差吸收到平均值之中；同时呈现最差地区误差、严重误差率及高分歧内容表现，才能较为完整地识别模型风险。

从理论与方法上看，这一设计进一步区分了样本内统计关联与跨地区预测效度。原 D3CODE 研究通过关联和中介分析发现，个体道德价值能够解释部分冒犯感知差异[16]；与之相对，跨地区实验把目标地区整体移出训练集，考察这一关系能否在新的群体环境中继续降低误差。前者呈现样本内关系，后者检验外部适用范围，二者共同说明变量关联强度与迁移能力需要分别检验。计算传播研究强调把传播现象转化为可计算问题，同时也要求区分预测与解释的认识任务[19-21]。沿着这一思路，样本外预测可以检验地区标签或价值量表是否达到预期的经验覆盖范围，分组误差则把总体指标还原到具体受众与内容情境。陈慧敏等将内容识别、美学计算和传播问题逐层衔接，展示了计算结果进入传播讨论的一种路径[24]；在此基础上，域外失败和误差分布被纳入分析，使模型表现既成为研究结果，也成为后续访谈、话语分析和制度研究的问题线索。

以上分析也为平台内容审核带来实践启示。内容审核通过分类与处置参与设定公共表达边界[7-8]，模型性能的改善也应当与误差分配和复核程序一并考察。平台可以把地区差异、议题差异和评分分歧纳入常规审计，在总体指标之外持续追踪最差群体表现，并将高分歧、低置信度和规则边界模糊的内容优先交由多元标注或人工复核。地区与价值信息更适合服务于群体层面的误差检查；若直接进入个人删除、降权或处罚规则，则可能把统计代理转化为群体画像，进而固化差别化处置。文化敏感的审核机制需要在模型训练、群体审计、人工复核和申诉反馈之间形成连续的评估过程，其有效性有赖于这一过程，而非字段数量本身。

不容回避的是，上述证据的适用范围仍受数据与测量条件限制。全部评论以英文呈现，完成任务所需的语言能力和语用经验可能进入地区差异；原始会话、互动关系和事件背景未

被保留，分析所观察的是任务情境中的再语境化判断。八个地区的分类相对粗略，尚难展开国家内部的城乡、族群、阶层和代际差别，道德价值问卷的跨地区测量等值性也有待专门检验。此外，0—4 五级评分属于顺序尺度，为便于模型比较，当前分析采用回归处理；评论与标注者之间又形成交叉重复结构，分组外测解决了数据泄漏问题，但地区、价值和 SHAP 结果仍属于预测关联。未来研究可结合有序模型和多层模型开展稳健性检验，并引入多语种材料、开放式判断理由、访谈和真实平台复核记录，进一步分析生产者意图、受众解释与平台规则如何共同塑造冒犯判断。

进一步而言，实验任务中的冒犯性评分与真实平台上的删除、降权、传播、申诉和用户信任处于不同分析层次。当前研究识别了感知预测的差异及其外推边界，治理后果还需在真实审核流程中加以验证。后续研究若能把模型误差、受众解释和复核结果连接起来，便可进一步判断分歧报告与多元复核能否改善程序公平，也能为统一规则与文化差异之间的协调提供更充分的经验证据。

参考文献

[1] DAVANI A, DÍAZ M, BAKER D, et al. D3CODE: Disentangling Disagreements in Data across Cultures on Offensiveness Detection and Evaluation[C]//Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. Miami, Florida, USA: Association for Computational Linguistics, 2024: 18511-18526.

[2] CULPEPER J. Impoliteness: Using Language to Cause Offence[M]. Cambridge: Cambridge University Press, 2011.

[3] HALL S. Encoding/Decoding[G]//HALL S, HOBSON D, LOWE A, et al. Culture, Media, Language: Working Papers in Cultural Studies, 1972–79. London: Hutchinson, 1980: 128-138.

[4] MORLEY D. The Nationwide Audience: Structure and Decoding[M]. London: British Film Institute, 1980.

[5] MARWICK A E, BOYD D. I Tweet Honestly, I Tweet Passionately: Twitter Users, Context Collapse, and the Imagined Audience[J]. New Media & Society, 2011, 13(1): 114-133.

[6] DAVIS J L, JURGENSON N. Context Collapse: Theorizing Context Collusions and Collisions[J].Information, Communication & Society, 2014, 17(4): 476-485.

[7] GILLESPIE T. Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media[M]. New Haven: Yale University Press, 2018.

[8] GORWA R, BINNS R, KATZENBACH C. Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance[J]. Big Data & Society, 2020, 7(1):1-15.

[9] 许向东，王怡溪. 智能传播中算法偏见的成因、影响与对策[J]. 国际新闻界, 2020, 42(10): 69-85.

[10] 史安斌，俞雅芸. “技术后冲”时代全球互联网治理的趋势与路径：基于五大主题的文献分析[J]. 新闻与传播评论, 2023, 76(3): 17-26.

[11] AROYO L, WELTY C. Truth Is a Lie: Crowd Truth and the Seven Myths of Human Annotation[J].AI Magazine, 2015, 36(1): 15-24.

[12] MOSTAFAZADEH DAVANI A, DÍAZ M, PRABHAKARAN V. Dealing with Disagreements: Looking Beyond the Majority Vote in Subjective Annotations[J]. Transactions of the Association for Computational

Linguistics, 2022, 10: 92-110.

[13] SAP M, SWAYAMDIPTA S, VIANNA L, et al. Annotators with Attitudes: How Annotator Beliefs and Identities Bias Toxic Language Detection[C]//Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Seattle, United States: Association for Computational Linguistics, 2022: 5884-5906.

[14] PEI J, JURGENS D. When Do Annotator Demographics Matter? Measuring the Influence of Annotator Demographics with the POPQUORN Dataset[C]//Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII). Toronto, Canada: Association for Computational Linguistics, 2023: 252-265.

[15] ATARI M, HAIDT J, GRAHAM J, et al. Morality beyond the WEIRD: How the Nomological Network of Morality Varies across Cultures[J]. Journal of Personality and Social Psychology, 2023, 125(5): 1157-1188.

[16] MOSTAFAZADEH DAVANI A, DÍAZ M, BAKER D, et al. Disentangling Perceptions of Offensiveness: Cultural and Moral Correlates[C]//FAccT '24: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency. New York, NY, USA: Association for Computing Machinery, 2024: 2007-2021.

[17] ZHOU L, KARAMOLEGKOU A, CHEN W, et al. Cultural Compass: Predicting Transfer Learning Success in Offensive Language Detection with Cultural Features[C]//Findings of the Association for Computational Linguistics: EMNLP 2023. Singapore: Association for Computational Linguistics, 2023: 12684-12702.

[18] ROBERTS D R, BAHN V, CIUTI S, et al. Cross-validation Strategies for Data with Temporal, Spatial, Hierarchical, or Phylogenetic Structure[J]. Ecography, 2017, 40(8): 913-929.

[19] 王成军. 计算传播学：作为计算社会科学的传播学[G]//巢乃鹏. 中国网络传播研究： 2014（第 8 辑）. 南京: 南京大学出版社, 2015: 193-208.

[20] 李雪莲, 刘德寰. 预测与解释：走向因果表征的传播学[J]. 国际新闻界, 2023, 45(7): 157-176.

[21] HOFMAN J M, WATTS D J, ATHEY S, et al. Integrating Explanation and Prediction in Computational Social Science[J]. Nature, 2021, 595(7866): 181-188.

[22] KE G, MENG Q, FINLEY T, et al. LightGBM: A Highly Efficient Gradient Boosting Decision Tree [C]//GUYON I, von LUXBURG U, BENGIO S, et al. Advances in Neural Information Processing Systems: vol. 30. Red Hook, NY, USA: Curran Associates, Inc., 2017: 3146-3154.

[23] LUNDBERG S M, LEE S I. A Unified Approach to Interpreting Model Predictions[C] GUYON I, VON LUXBURG U, BENGIO S, et al. Advances in Neural Information Processing Systems: vol. 30. Red Hook, NY, USA: Curran Associates, Inc., 2017: 4765-4774.

[24] 陈慧敏, 陈祉伊, 吴宜檬, 金兼斌. 生成式人工智能介入下国家形象的视觉框架及其效果——基于推特社交媒体平台中国图像数据的实证分析[J]. 当代传播, 2025(3): 41-48.

Audience Differences in Cross-Cultural Judgments of Offensiveness and the Limits of Unified Content Moderation Models: An Interpretable Machine Learning Study of D3CODE

CHEN Yilai

(Yangming College, Ningbo University, Ningbo / Zhejiang 315211)

Abstract: Global platforms often rely on unified models to moderate content for culturally diverse audiences, yet the same expression may receive sharply different offensiveness ratings across regions and individuals. Existing research emphasizes textual features, paying less attention to regional background, individual values, and model performance in unseen regions. Drawing on 4,554 English comments, 4,309 annotators, and 144,730 valid ratings in D3CODE, this study uses ablation analysis, leave-one-region-out validation, error slicing, and SHAP to examine how text, region, and moral values contribute to predicting offensiveness ratings. Regional information yields greater predictive gains than individual moral values. The model combining region, values, and prespecified interactions performs best, but individual values do not reduce macro-averaged error across the eight regions. Prediction error rises with rating disagreement and is higher for content concerning religion, social groups, and gender and relationship norms. Offensiveness judgments therefore cannot be inferred from text alone, and unified moderation models do not fully capture cross-cultural audience variation. Platforms should report performance separately by region, topic, and level of disagreement, alongside aggregate metrics.

Keywords: cross-cultural communication; offensiveness judgments; audience variation; platform governance; content moderation; computational communication

作者简介（可选）：陈祎来（fish@ooo.cat），宁波大学阳明创新班本科生。本项目代码仓库：https://github.com/tallusil/cross-cultural-offensive。