毒性检测模型增强

Enhancement of Toxicity Detection Model

Chen Huihua (ASCII LAB) 2024/10/11

免责声明：PPT内例子都是为了理解语境

毒性检测模型

毒性定义：

负面内容，包括仇恨言论、攻击性、辱骂性或亵渎性语言、攻击性、不文明、侮辱、骚扰、威胁、网络欺凌、讽刺、讽刺和贬损性语言

毒性危害：

- 影响使用者的使用体验，造成提前离场
- 危害少年身心健康，网络安全、社会和谐

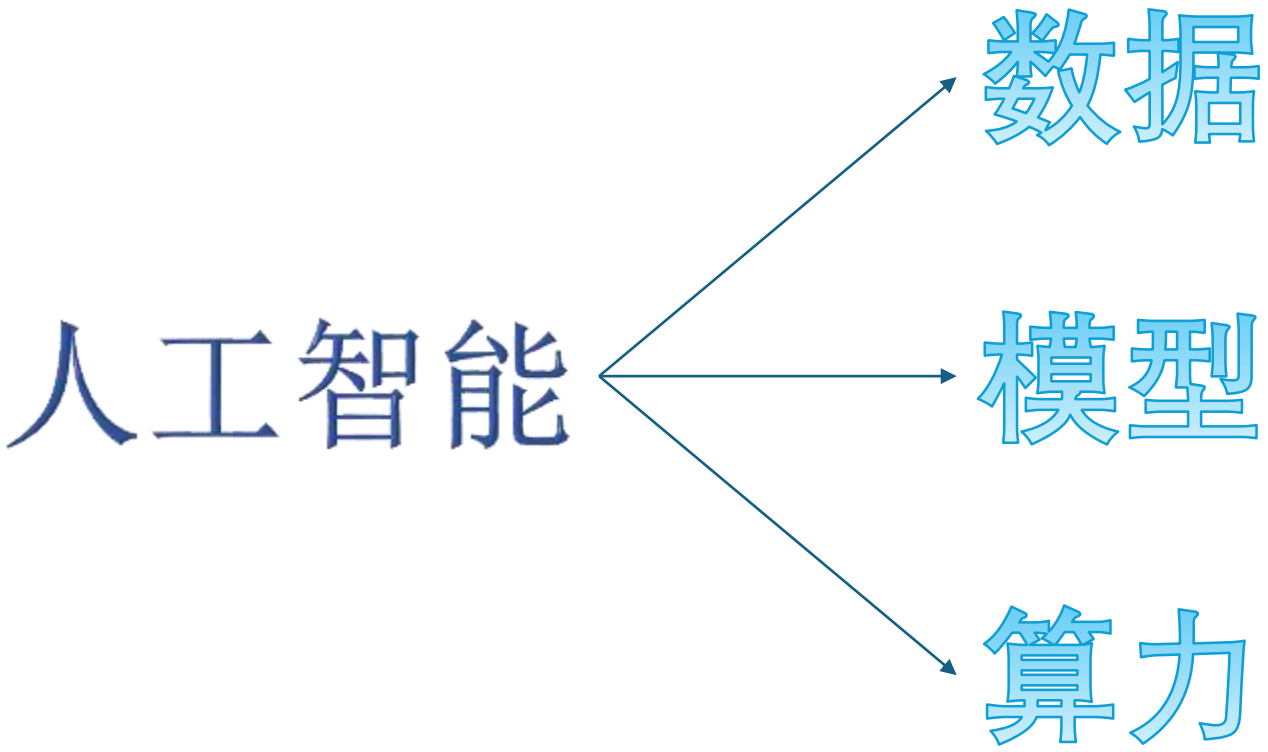
Reference: A Systematic Review of Toxicity in Large Language Models: De?nitions, Datasets, Detectors, Detoxification Methods and Challenges, 2024年7月

毒性检测模型增强

毒性检测任务复杂性：

- 同一句话，对于不同的人毒性不同
- 同一句话，在不同场景下毒性不同

传统毒性检测 ↛ 现实需求 → 毒性检测增强


人工智能
数据
模型
算力

数据

拿来就能用的数据集

可靠性

研究集中在人机交互领域
注释分歧并非仅存在于毒性检测

人工注释分歧

注释者的文化背景、生活经历
语或互动对注释的毒性不同

模型注释能力

使用模型自动标注毒性的能力

完整性

多语言多模态

经典国骂，广泛流传
语音骂人，防不胜防

特定领域泛化

通用模型泛化到特定领域
游戏：三国X，狼人X

模型

提出一个方法

准确性

隐形毒性检测

隐晦地骂，你得检测到
“啊对对对，你真的太聪明了”

鲁棒性

对抗毒性检测

故意骂错，你得检测到
“fuck” → “fμck”

偏见

毒性检测偏见

考虑模型使用人群
模型仅使用男性数据训练，
只针对女性地毒性话语不一定有毒

数据

1. 数据集可靠性
   - 人工注释分歧
   - 自动注释能力
2. 数据集完整性
   - 多语言多模态
   - 特定领域泛化

模型

1. 准确性
   - 隐性毒性检测
2. 鲁棒性
   - 对抗毒性检测
3. 偏见
   - 毒性检测偏见

目录

数据

1. 数据集可靠性
   - 人工注释分歧
   - 自动注释能力
2. 数据集完整性
   - 多语言多模态
   - 特定领域泛化

模型

1. 准确性
   - 隐性毒性检测
2. 鲁棒性
   - 对抗毒性检测
3. 偏见
   - 毒性检测偏见

01. 数据

- 1.1. MuTox: Universal MUltilingual Audio-based TOXicity Dataset and Zero-shot Detector, ACL2024

研究动机：

基于音频的毒性检测研究有限，尤其是非英语语种

研究思路：

目标： 推动多语言音频毒性检测的研究

做法： 开发数据集，要求：毒性、多语言、音频

MuTox: Universal Multilingual Audio-based Toxicity Dataset and Zero-shot Detector, ACL2024

构建数据集

1.选择数据源

- **多语种：Common Voice —** 覆盖多种语言的大规模语音数据集
- **真实性：Seamless Align Expressive —** 网络收集的语音识别数据集

2.数据预处理

- **筛选：** 基于语义获取和认知负荷，选择2-8秒之间的语音
- **ASR：** 使用自动语音识别系统将音频转换为文本

3. 毒性标注

- **模型标注：** 使用现有的文本毒性检测模型标注数据毒性
- **人工标注：** 组织专业团队根据注释指南进行数据注释

MuTox: Universal Multilingual Audio-based Toxicity Dataset and Zero-shot Detector, ACL2024

构建数据集 – 注释指南

语句

- 语义：明确的词句含义
- 言外性：语句对于听者的影响
- 例子：如果你的家人有什么问题，那就太遗憾了

注释

- **1.** 语句是否含有毒性
- **2.** 如果含有毒性，你认为是什么原因？
  - **A)** 具体的词汇或短语？请指明
  - **B)** 某种言外性？
    - 咄咄逼人的声音？咄咄逼人的语气？隐晦的威胁？

MuTox: Universal Multilingual Audio-based Toxicity Dataset and Zero-shot Detector, ACL2024

构建数据集

1.选择数据源

- **Common Voice：** 覆盖多种语言的大规模语音数据集
- **Seamless Align Expressive：** 网络收集的语音识别数据集

2.数据预处理

- **筛选：** 基于语义获取和认知负荷，选择2-8秒之间的语音
- **ASR：** 使用自动语音识别系统将音频转换为文本

3. 毒性 标注

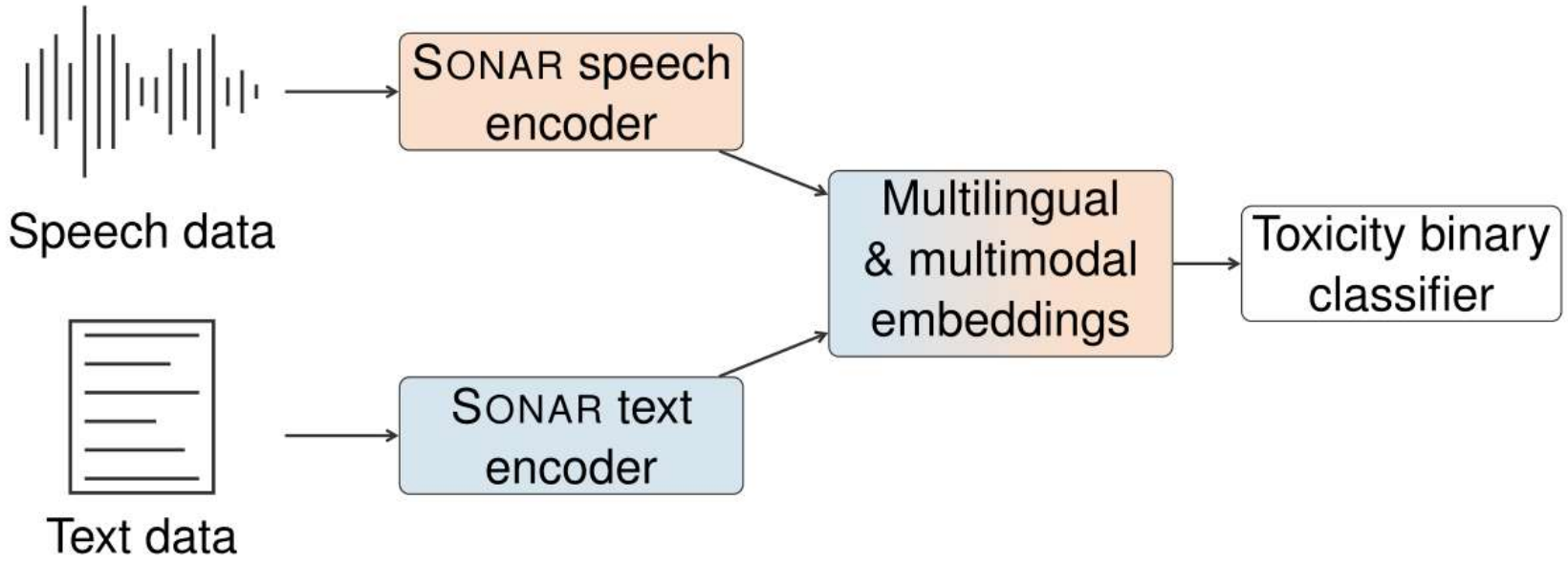
- **模型标注：** 使用现有的毒性检测模型标注数据毒性
- **人工标注：** 组织专业团队根据注释指南进行数据注释

4.数据集平衡

- **类别平衡：** 确保数据集中各类别毒性(侮辱、仇恨)数量平衡
- **语言平衡：** 确保数据集中不同语言数量平衡

MuTox: Universal Multilingual Audio-based Toxicity Dataset and Zero-shot Detector, ACL2024

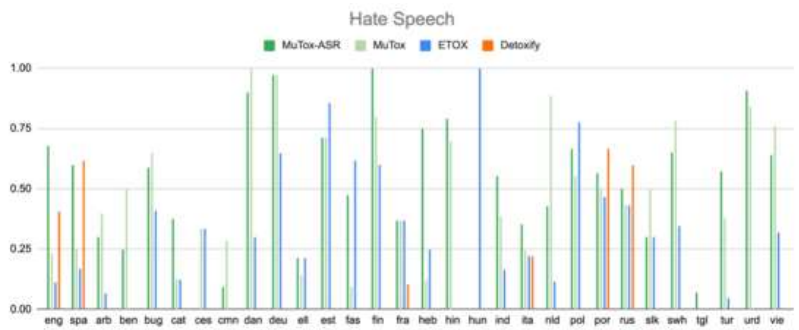
毒性检测模型

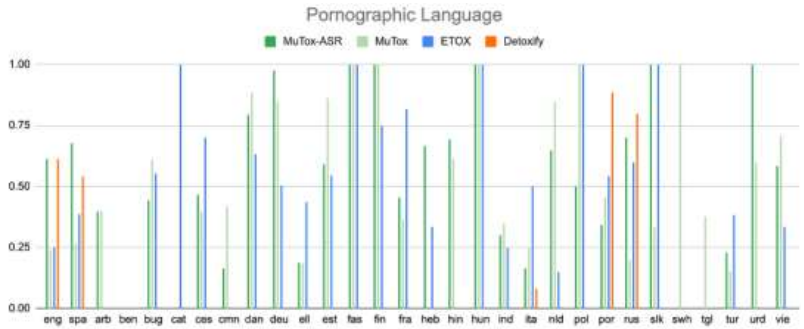


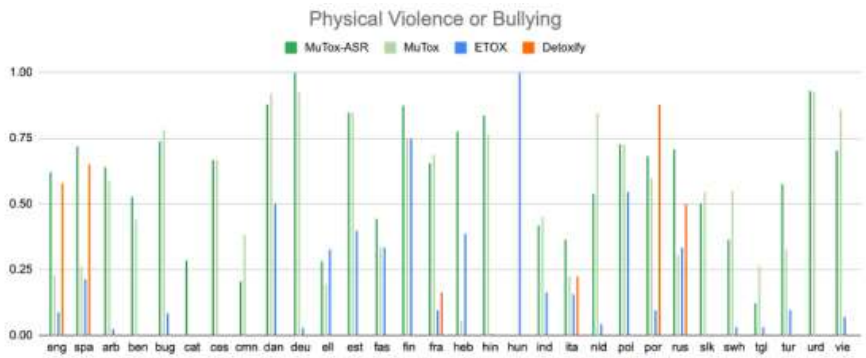
多模态	多模态	三层全连接的分类器
编码器	嵌入	**(1024, 512, 128)**

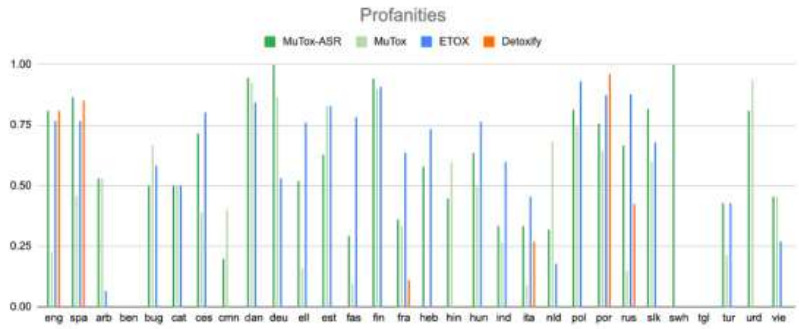
MuTox: Universal Multilingual Audio-based Toxicity Dataset and Zero-shot Detector, ACL2024

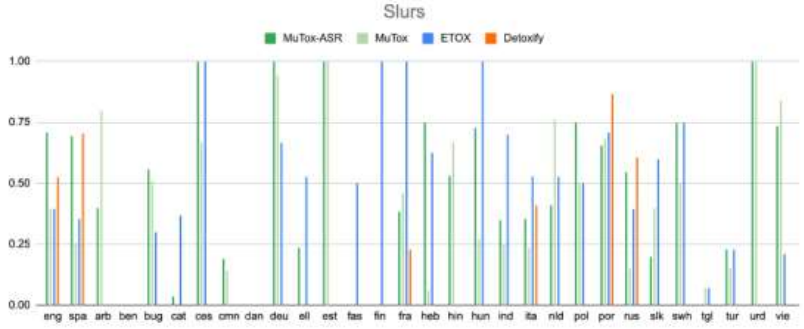
实验结果











对比表现最好的音频毒性检测器，提供了十倍以上的语言覆盖范围
对比覆盖范围最广的音频毒性检测器，F1-score实现了100%的增长

毒性检测 - 多模态多语言

本篇总结

作者提出了一个音频毒性检测的数据集

并通过简单的网络架构就达到了不错的性能（覆盖范围和检测效果）

方向扩展

数据角度：音频视频数据

模型角度：更复杂的网络架构，引入注意力机制、多模态联合分析

MuTox: Universal Multilingual Audio-based Toxicity Dataset and Zero-shot Detector, ACL2024

数据

1. 数据集可靠性
   - 人工注释分歧
   - 自动注释能力
2. 数据集完整性
   - 多语言多模态
   - 特定领域泛化

模型

1. 准确性
   - 隐性毒性检测
2. 鲁棒性
   - 对抗毒性检测
3. 偏见
   - 毒性检测偏见

01. 数据

- 1.2. Automated Identification of Toxic Code Reviews Using ToxiCR, TOSEM 2023

研究动机：

- 开源社区中有毒文本泛化，导致许多开发者离开开源社区
- 传统毒性检测器不能在软件开发领域中直接使用

例子：Kill 进程，进程的状态为 Dead，URL

研究思路：

目标： 如何将通用毒性检测模型泛化到软件工程领域

做法： 模型架构（增加模块）、推理阶段（prompt）、**训练数据**

构建数据集

1.数据挖掘

- 挑选流行的开源软件项目(Android、Chromium OS、OpenStack和LibreOffice)，爬取代码审查评论

2.数据筛选

- 查找关键词“bot”等，排除所有机器人账号的评论
- 分层抽样：有害评论更罕见，确保有毒和无毒评论的数量平衡

3.数据标注

- 模型标注：使用现有的毒性检测模型标注数据毒性
- 人工标注：组织专业团队根据注释指南进行数据注释

MuTox: Universal Multilingual Audio-based Toxicity Dataset and Zero-shot Detector, ACL2024

数据预处理

1. 强制性预处理步骤

- URL移除 [正则表达式]
- 缩写扩展：将缩写词扩展为完整形式 → “TMD” [词典：153个缩写]
- 符号移除：移除特殊符号 → “fu*ck” [正则表达式]
- 重复消除：修正故意拼写错误的词汇 → “fuuck”
- 对抗识别：识别并更正非常规拼写代替的毒性词 → “nimade”

2. 可选预处理步骤

- 标识符拆分：“F_u_c_k = 1” [正则表达式]
- 编程关键字移除：移除“kill”等 [关键字列表]
- 计数毒性词：计算文本中毒性词的出现次数 [毒性词词典]

MuTox: Universal Multilingual Audio-based Toxicity Dataset and Zero-shot Detector, ACL2024

实验分析

文本数量：19651

Group	Algo	Vectorizer	Preprocessing			Non-toxic			Toxic			A
			profane-count	kwrd-remove	id-split	P_0	R_0	$F1_0$	P_1	R_1	$F1_1$	
CLE	DT	tfidf	✓	✓	-	0.960	0.968	0.964	0.862	0.830	0.845	0.942
	GBT	tfidf	✓	✓	-	0.938	0.981	0.959	0.901	0.729	0.806	0.932
	LR	tfidf	✓	✓	-	0.932	0.981	0.956	0.898	0.698	0.785	0.927
	RF	tfidf	✓	-	-	0.964	**0.981**	0.972	**0.917**	0.845	0.879	0.955
	SVM	tfidf	✓	✓	-	0.939	0.977	0.958	0.886	0.736	0.804	0.931
DNN	DPCNN	fasttext	✓	-	-	0.964	0.973	0.968	0.889	0.846	0.863	0.948
	LSTM	glve	✓	✓	✓	0.944	0.974	0.959	0.878	0.756	0.810	0.932
	BiLSTM	fasttext	✓	-	✓	0.966	0.975	0.971	0.892	0.858	0.875	0.953
	BiGRU	glove	✓	-	✓	0.966	0.976	0.971	0.897	0.856	0.876	0.954
Transormer	BERT	bert	-	✓	-	**0.970**	0.978	**0.974**	0.907	**0.874**	**0.889**	**0.958**
STRUDEL	文本数量：611										0.4	0.83
Sarker	文本数量：10673										0.87	0.92

实验分析

Models	Non-toxic			Toxic			Accuracy
	P_0	R_0	$F1_0$	P_1	R_1	$F1_1$	
Perspective API [7] (off-the-shelf)	0.92	0.79	0.85	0.45	0.70	0.55	0.78
Strudel Tool (off-the-shelf) [73]	0.93	0.76	0.83	0.43	0.77	0.55	0.76
Strudel (retrain) [78]	**0.97**	**0.96**	**0.97**	**0.85**	**0.86**	**0.85**	**0.94**
DPCNN (retrain) [77]	0.94	0.95	0.94	0.81	0.76	0.78	0.91

Data is all you need !

实验分析

计数亵渎词、编程关键词去除、标识符切割

Group	Algo	Vectorizer	Preprocessing			Non-toxic			Toxic			A
			profane-count	kwrd-remove	id-split	P_0	R_0	$F1_0$	P_1	R_1	$F1_1$	
CLE	DT	tfidf	✓	✓	-	0.960	0.968	0.964	0.862	0.830	0.845	0.942
	GBT	tfidf	✓	✓	-	0.938	0.981	0.959	0.901	0.729	0.806	0.932
	LR	tfidf	✓	✓	-	0.932	0.981	0.956	0.898	0.698	0.785	0.927
	RF	tfidf	✓	-	-	0.964	**0.981**	0.972	**0.917**	0.845	0.879	0.955
	SVM	tfidf	✓	✓	-	0.939	0.977	0.958	0.886	0.736	0.804	0.931
DNN	DPCNN	fasttext	✓	-	-	0.964	0.973	0.968	0.889	0.846	0.863	0.948
	LSTM	glve	✓	✓	✓	0.944	0.974	0.959	0.878	0.756	0.810	0.932
	BiLSTM	fasttext	✓	-	✓	0.966	0.975	0.971	0.892	0.858	0.875	0.953
	BiGRU	glove	✓	-	✓	0.966	0.976	0.971	0.897	0.856	0.876	0.954
Transormer	BERT	bert	-	✓	-	**0.970**	0.978	**0.974**	0.907	**0.874**	**0.889**	**0.958**

对于可选预处理步骤，每个模型的最佳性能配置

毒性检测 – 特定领域泛化

本篇总结

特定于软件工程领域，本文扩展了数据集、提出了针对性预处理操作

实验体现了数据量和预处理步骤对性能的帮助

方向扩展

为了能将毒性检测泛化到各个领域

数据：数据生成技术、针对不同领域的数据预处理步骤

MuTox: Universal Multilingual Audio-based Toxicity Dataset and Zero-shot Detector, ACL2024

数据

1. 数据集可靠性
   - 人工注释分歧
   - 自动注释能力
2. 数据集完整性
   - 多语言多模态
   - 特定领域泛化

模型

1. 准确性
   - 隐性毒性检测
2. 鲁棒性
   - 对抗毒性检测
3. 偏见
   - 毒性检测偏见

02. 模型

- 2.1. Tox-BART: Leveraging Toxicity Attributes for Explanation Generation of Implicit Hate Speech, ACL2024

介绍

为什么警察杀黑人不入狱?

因为虐待动物不会被判刑。

Why do police officers never get incarcerated for shooting unarmed black men? Because animal cruelty is not punishable by jail time.

Implied Stereotype Generation

black people are animals

黑人是动物

0.31 0.07

Toxicity Prediction API

毒性检测API

Target Prediction API

毒性目标检测API

"Police officers"

0.97 0.98

Toxicity Prediction API

Target Prediction API

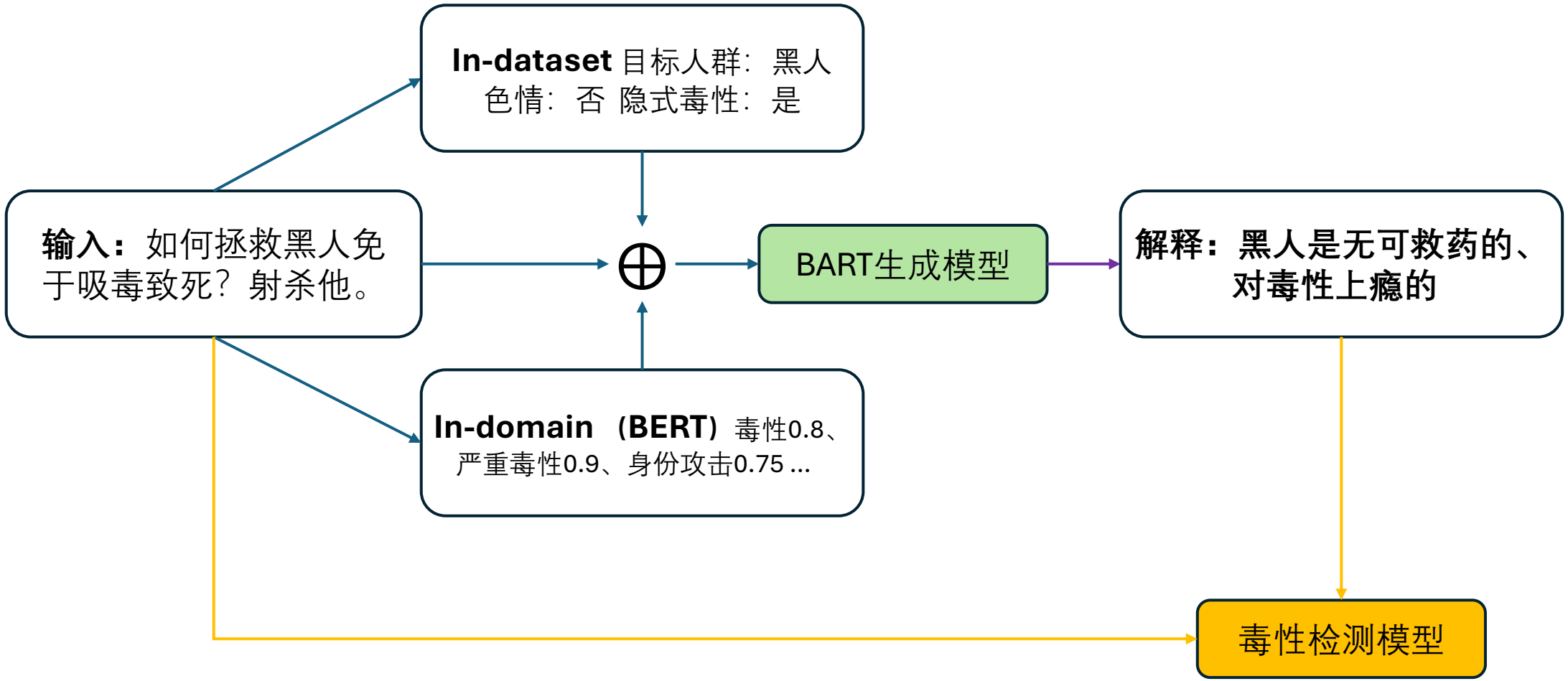
"Black men"

隐性毒性言论

1. 是什么
2. 为什么
3. 怎么做

Tox-BART: Leveraging Toxicity Attributes for Explanation Generation of Implicit Hate Speech, ACL2024

注入知识信号



Tox-BART: Leveraging Toxicity Attributes for Explanation Generation of Implicit Hate Speech, ACL2024

注入知识信号

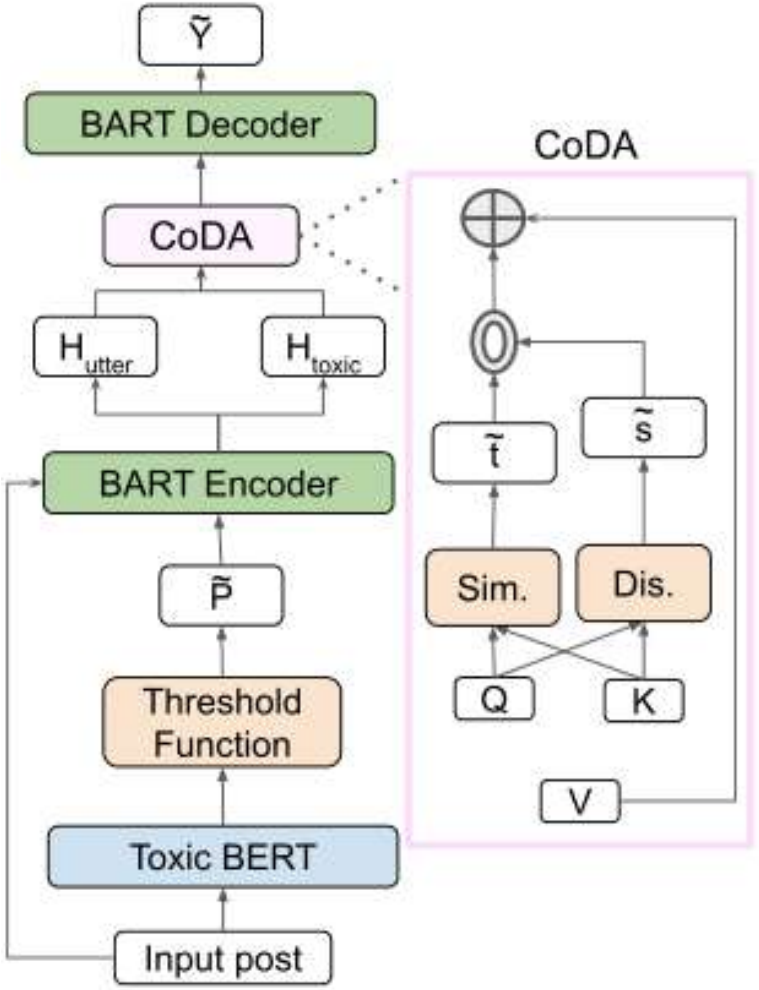
输入：如何拯救黑人免于吸毒致死？射杀他。

In-dataset 目标人群：黑人 色情：否 隐式毒性：是

⊕

BART生成模型

In-domain （BERT）毒性0.8、严重毒性0.9、身份攻击0.75 ...



Tox-BART: Leveraging Toxicity Attributes for Explanation Generation of Implicit Hate Speech, ACL2024

性能分析

自动评估

Method	数据集	SBIC			LatentHatred		
	指标	BLEU	ROUGE-L	BERTScore	B	R	BS
GPT-2		62.72	62.72	59.04	30.94	21.99	82.71
BART		**72.17**	**70.83**	78.05	38.38	17.65	90.37
MIXGEN - *Imp*		72.12	69.84	**80.91**	46.28	35.78	92.09
MIXGEN - *Exp*		68.41	66.40	80.37	47.23	**36.26**	92.12
MIXGEN - *Exp* + *Imp*		70.27	67.69	80.23	47.00	33.09	90.8
Tox-BART$_{C1}$	In-domain	64.89	63.83	64.52	41.94	26.28	89.47
Tox-BART$_{C2}$	In-dataset	69.85	68.23	75.78	**47.72**	34.70	**92.89**
GPT-3.5 (Zeroshot)		37.45	15.36	**90.10**	33.57	10.40	90.06

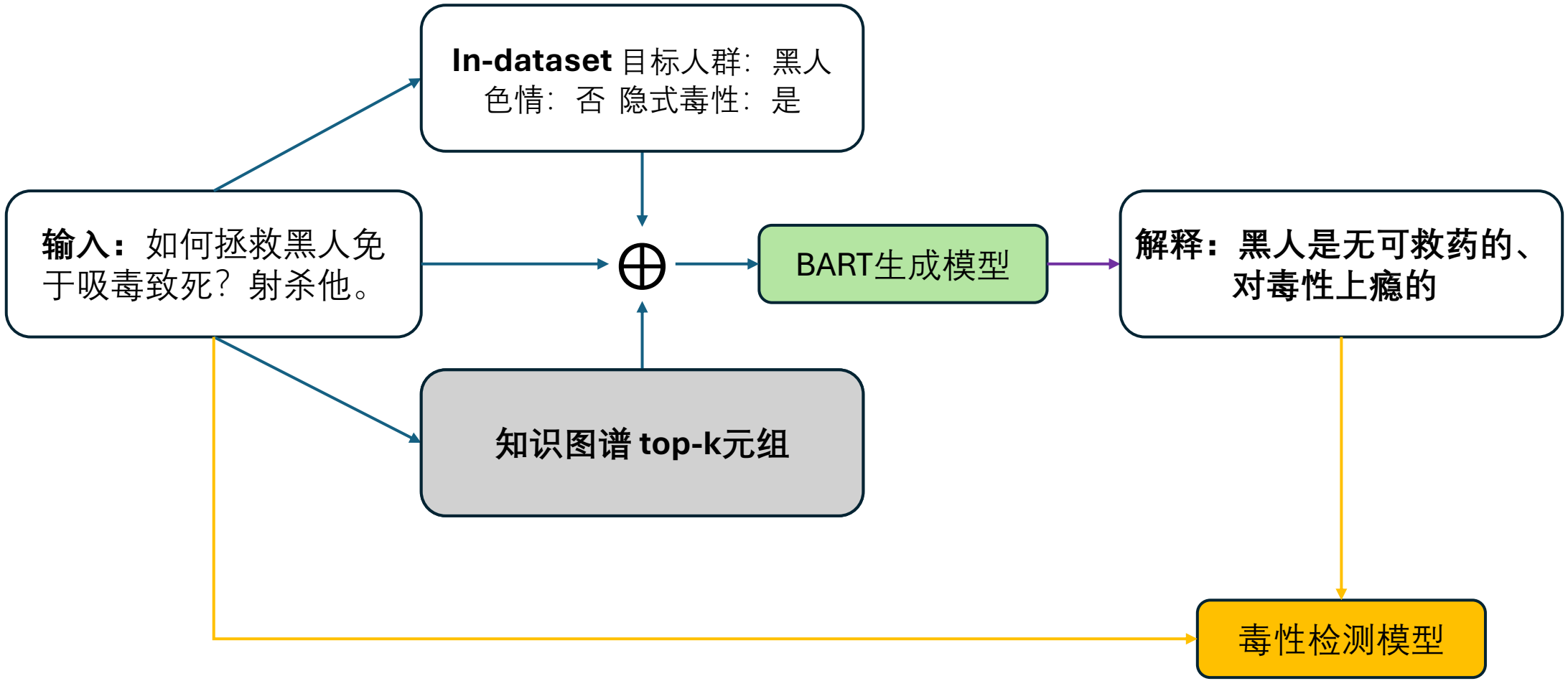
人工评估

Method	Flu.	Coh.	Spe.	Sim.	Tar.
Tox-BART$_{C1}$	4.52 (±0.76)	3.95 (±0.99)	3.67 (±0.92)	3.47 (±1.00)	0.78 (±0.29)
GPT-3.5	4.17 (±0.9)	3.74 (±0.92)	3.27 (±1.07)	2.78 (±1.14)	0.49 (±0.4)

Method	Toxicity↑
Tox-BART$_{C1}$	0.89 (±0.21)
GPT-3.5	0.33 (±0.32)

Tox-BART: Leveraging Toxicity Attributes for Explanation Generation of Implicit Hate Speech, ACL2024

审查KG元组质量



Tox-BART: Leveraging Toxicity Attributes for Explanation Generation of Implicit Hate Speech, ACL2024

审查KG元组质量

假设：添加 top-k 元组有助于提高模型的生成能力，
那么当 top-k 损坏时，生成应该会恶化

Method	知识图谱	ConceptNet						StereoKG					
	隐毒数据集	SBIC			LatentHatred			SBIC			LatentHatred		
	指标	B	R	BS	B	R	BS	B	R	BS	B	R	BS
BART Baseline		72.17	70.83	78.05	38.38	17.65	90.37	72.17	70.83	78.05	38.38	17.65	90.37
Top-k		68.41	66.4	80.37	47.23	36.26	92.12	63.57	61.30	76.39	46.39	35.37	92.03
Bottom-k		68.97	66.80	80.95	47.40	35.90	92.15	60.31	58.09	73.44	46.92	35.94	92.04
Random-k		69.69	67.47	81.63	48.34	37.18	92.31	60.80	58.45	73.87	47.27	36.12	92.07

KG 总结 KG 注入，在目前的一些主观的推理任务中，例如隐性毒性解释，并没有完全发挥作用

提高能力：特定领域的知识图谱检索方法、知识图谱元组排序方法，提升top-k元组的质量

Tox-BART: Leveraging Toxicity Attributes for Explanation Generation of Implicit Hate Speech, ACL2024

毒性检测 – 隐形毒性检测

本篇总结

- 分析了使用**indataset**标注数据和机器标注数据，对模型性能的影响
- 指出了使用知识图谱进行隐性毒性检测的不足之处
- 并且通过模型生成解释，成功地提高了隐性毒性检测的成功率

思路扩展

减轻毒性的主观任务无法完全自动化

未来的工作可以聚焦**KG**的进一步开发

MuTox: Universal Multilingual Audio-based Toxicity Dataset and Zero-shot Detector, ACL2024

数据

1. 数据集可靠性
   - 人工注释分歧
   - 自动注释能力
2. 数据集完整性
   - 多语言多模态
   - 特定领域泛化

模型

1. 准确性
   - 隐性毒性检测
2. 鲁棒性
   - 对抗毒性检测
3. 偏见
   - 毒性检测偏见

02. 模型

- **2.2. MTTM: Metamorphic Testing for Textual Content Moderation Software, ICSE 2023**

MTTM 测试框架思路

总结变异关系

- 我们通过分析2000条真实用户的文本消息，总结了恶意用户规避审核的11种变异关系

选择目标词汇

- 使用TF-IDF算法选择目标词汇（有毒数据集经常出现、无毒数据集不常出现的词汇）

生成测试用例

- 对选中的目标词应用变异关系，生成测试用例

模型测试评估

- 将生成的测试用例输入到模型，判断是否能够被正确分类

重训改进模型

- 使用MTTM发现的错误案例，重新训练毒性检测模型，提高模型的鲁棒性

MTTM 测试框架实现

变异关系 & 生成测试用例

英语：基于字母　　汉语：基于象形

Perturbation Level		Perturbation Method	Examples in English	Examples in Chinese	Percentage
Character Level	1-1	Visual-based Substitution	a → α; C → (; l → 1	日 → 曰; 北 → 兆	12.3%
	1-2	Visual-based Splitting	K → |<; W → VV	好的 → 女子白勺	5.0%
	1-3	Visual-based Combination	Earn → Eam	不用 → 甭	0.8%
	1-4	Noise Injection	Hello → H**elll*o	致电 → 致*电	13.2%
	1-5	Char Masking	Hello → H*llo	新年快乐 → 新年快*	7.4%
	1-6	Character Swap	Weather → Waether	简单来说 → 简来单说	4.1%
Word Level	2-1	Language Switch	Hello → Hola; + → Add	龙 → 龍	14.9%
	2-2	Homophone Substitution	Die → Dye; Night → Nite	好吧 → 猴八; 这样 → 酱	36.4%
	2-3	Abbreviation Substitution	As Soon As Possible → ASAP	永远的神 → yyds	15.7%
	2-4	Word Splitting	Hello → Hell o	使用户满意 → 使用..户满意	6.6%
Sentence Level	3	Benign Context Camouflage	Golden State Warriors guard won't play Sunday, <add a spam sentence here>, due to knee soreness.	金融业增加值超香港, <在这里添加一条广告>, 是金融市场体系最完备、集中度最高的区域。	2.5%

MTTM 测试框架实验

生成测试用例 → 模型测试 → 重训改进模型

1. MTTM 框架生成的用例，有毒且真实
2. MTTM 框架生成的用例，是否会被审核软件误判?
3. 使用 MTTM 框架生成的用例，可以提高审核软件的性能

MTTM 测试框架实验

1. **MTTM** 框架生成的用例，有毒且真实
2. **MTTM** 框架生成的用例，是否会被审核软件误判?
3. 使用 **MTTM** 框架生成的用例，可以提高审核软件的性能

Level	Perturbation Methods	Abuse Detection				Spam Detection			Pron Detection		
		Google	Baidu	Huawei	AM	Baidu	Huawei	AM	Google	Baidu	Huawei
Char	Visual-based Substitution	19.4	28.0	75.9	91.2	51.0	75.7	84.0	36.9	35.2	47.2
	Visual-based Split	30.9	16.3	52.7	53.1	49.3	81.3	82.2	51.6	19.7	31.0
	Noise Injection (non-lang)	57.1	0.0	2.2	88.9	0.0	1.8	28.8	9.2	0.0	0.4
	Noise Injection (lang)	72.7	12.1	56.2	88.9	49.3	63.5	79.2	19.5	19.7	49.3
	Char Masking	50.8	19.8	50.3	88.9	47.2	58.1	78.9	10.7	38.0	47.9
	Char Swap	64.3	10.2	54.8	66.2	47.5	55.6	75.7	23.0	18.1	46.5
Word	Language Switch	57.7	38.0	76.3	84.1	35.7	49.3	53.9	32.7	39.4	49.3
	Homophone Substitution	73.4	26.8	77.4	85.6	48.9	75.7	77.1	22.6	36.6	47.2
	Abbreviation Substitution	83.9	22.7	63.4	88.9	52.2	82.5	83.6	32.1	38.0	48.6
	Visual Split	68.2	0.0	0.0	85.6	0.0	0.0	87.0	8.3	0.0	0.0
Sentence	Benign Context Camouflage	41.7	24.7	0.0	4.6	8.5	0.0	0.0	50.0	42.4	0.0
Multi	Perturbation Combinations	75.1	30.5	79.8	90.3	50.2	76.4	80.1	66.4	45.1	48.9

错误率

MTTM 测试框架实验

1. MTTM 框架生成的用例，有毒且真实
2. MTTM 框架生成的用例会被审核软件误判
3. 使用 MTTM 框架生成的用例，是否可以提高审核软件的性能？

Level	Perturb Methods	Ori	Aug
Char	Visual-Based Substitution	71.3	0.0
	Visual-Based Splitting	49.5	1.4
	Noise Injection (non-lang)	56.1	2.5
	Noise Injection (lang)	56.1	2.5
	Char Masking	43.9	2.5
	Char Swap	45.6	3.0
Word	Language Switch	76.2	5.9
	Homophone Substitution	62.5	3.1
	Abbreviation Substitution	76.2	2.2
	Visual Splitting	71.3	2.0
Sentence	Benign Context Camouflage	12.0	0.0
Multi	Perturbation Combinations	81.4	3.5

错误率

MTTM 测试框架实验

1. MTTM 框架生成的用例，有毒且真实
2. MTTM 框架生成的用例会被审核软件误判
3. 使用 MTTM 框架生成的用例，可以提高审核软件的性能

毒性检测 – 对抗毒性检测

本篇总结

提出框架，生成有毒用例，训练后可以提高检测模型的鲁棒性

论文优势

1. 区别对抗生成：有11种变异关系
2. 有两种语言设置（中英），并且能够再度推广到其他语言
3. 能够增强检测模型的鲁棒性，而不是单纯攻击

思路扩展

更多语言、自动化捕捉变异关系、实时训练

MuTox: Universal Multilingual Audio-based Toxicity Dataset and Zero-shot Detector, ACL2024

数据

1. 数据集可靠性
   - 人工注释分歧
   - 自动注释能力
2. 数据集完整性
   - 多语言多模态
   - 特定领域泛化

模型

1. 准确性
   - 隐性毒性检测
2. 鲁棒性
   - 对抗毒性检测
3. 偏见
   - 毒性检测偏见

02. 模型

- 2.3. Same Same, But Different: Conditional Multi-Task Learning for Demographic-Specific Toxicity Detection, WWW2023

毒性检测模型偏见

模型偏见：

- 针对不同群体的有毒言论可能存在显著差异
- 数据集中不同群体的分布可能不均衡

Same Same, But Different: Conditional Multi-Task Learning for Demographic-Specific Toxicity Detection, WWW2023

毒性检测模型偏见

模型偏见:

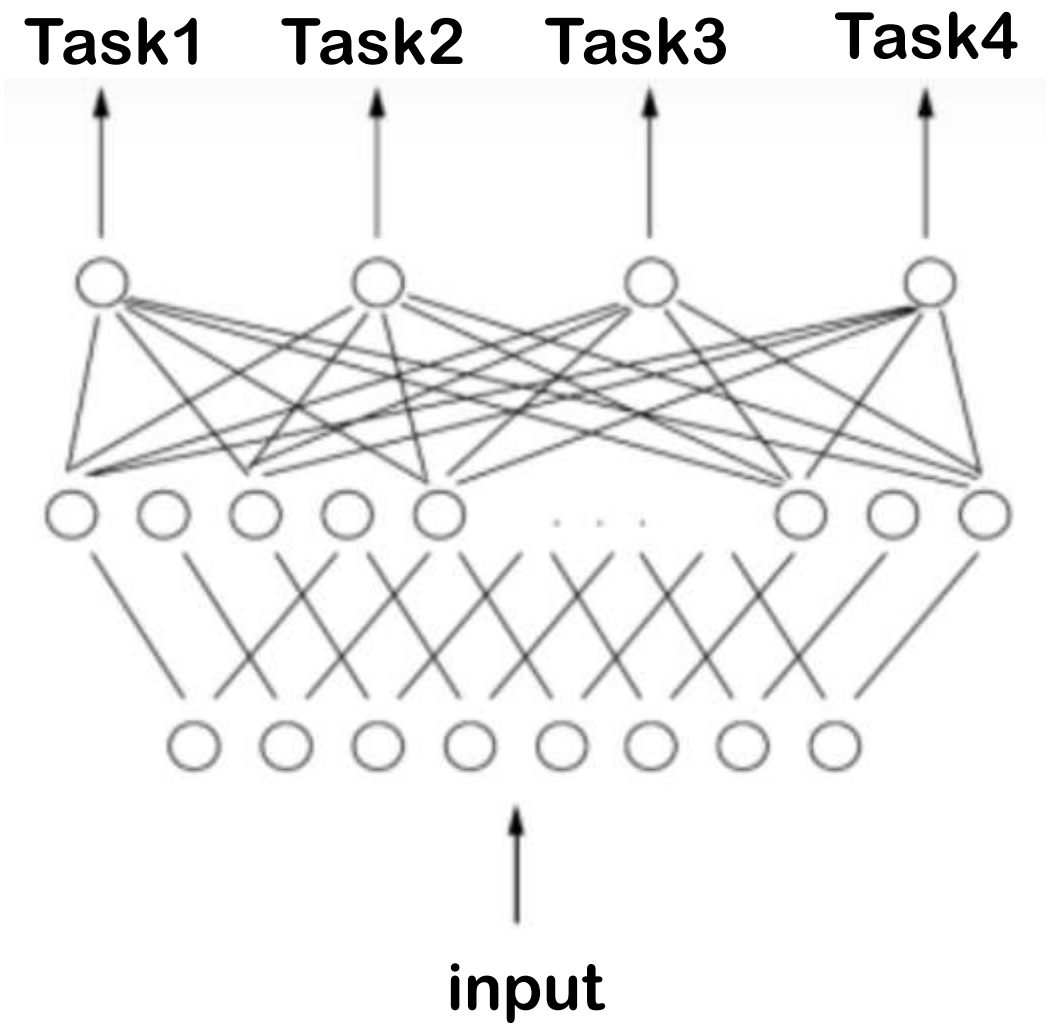
- 针对不同群体的有毒言论可能存在显著差异
- 数据集中不同群体的分布可能不均衡

求同存异:

- 学习到更广泛的毒性模式(不应该各自建立模型)
- 尊重少数群体的意见(不应该使用唯一模型)

Same Same, But Different: Conditional Multi-Task Learning for Demographic-Specific Toxicity Detection, WWW2023

传统多任务学习



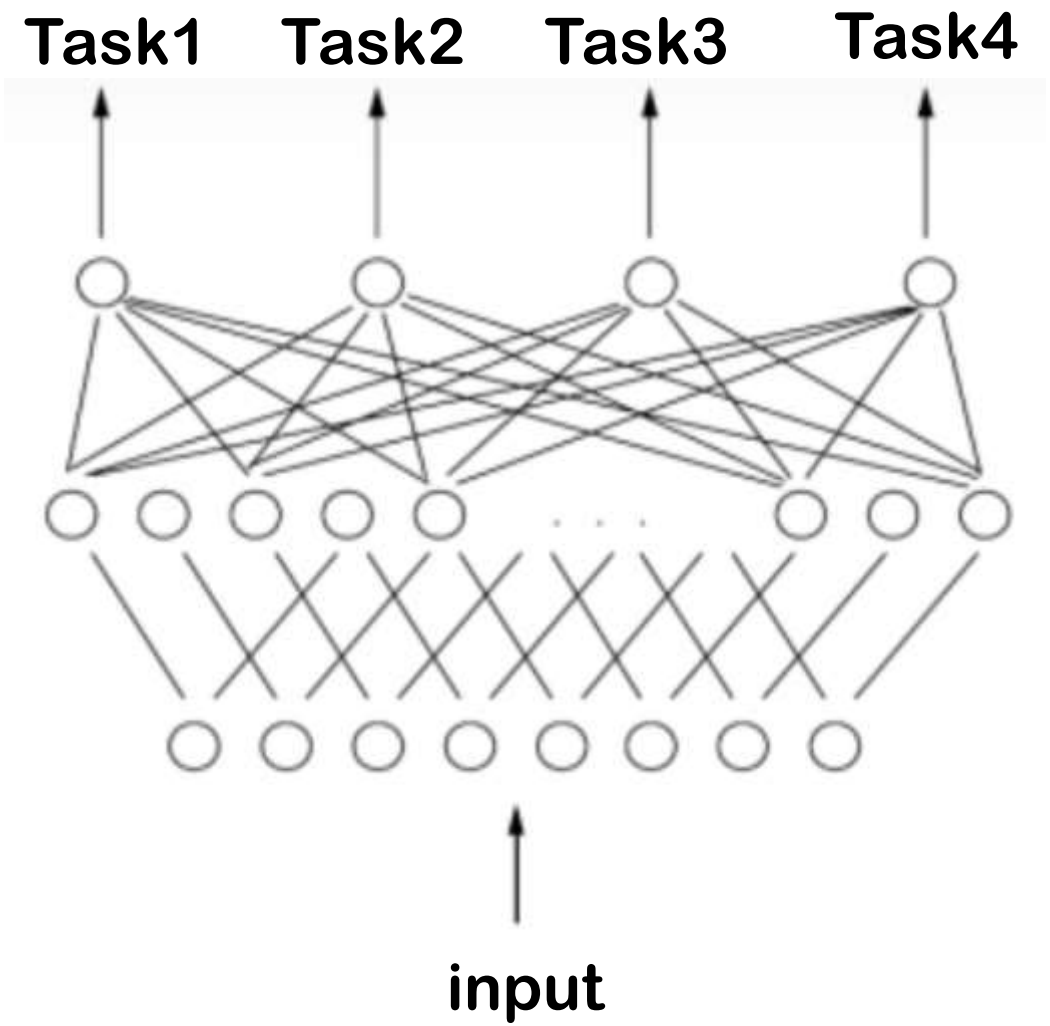
Task：针对不同群体的毒性检测

优点：求同存异

- 利用毒性语言间的共同模式
- 处理针对不同群体的毒性语言

Same Same, But Different: Conditional Multi-Task Learning for Demographic-Specific Toxicity Detection, WWW2023

传统多任务学习



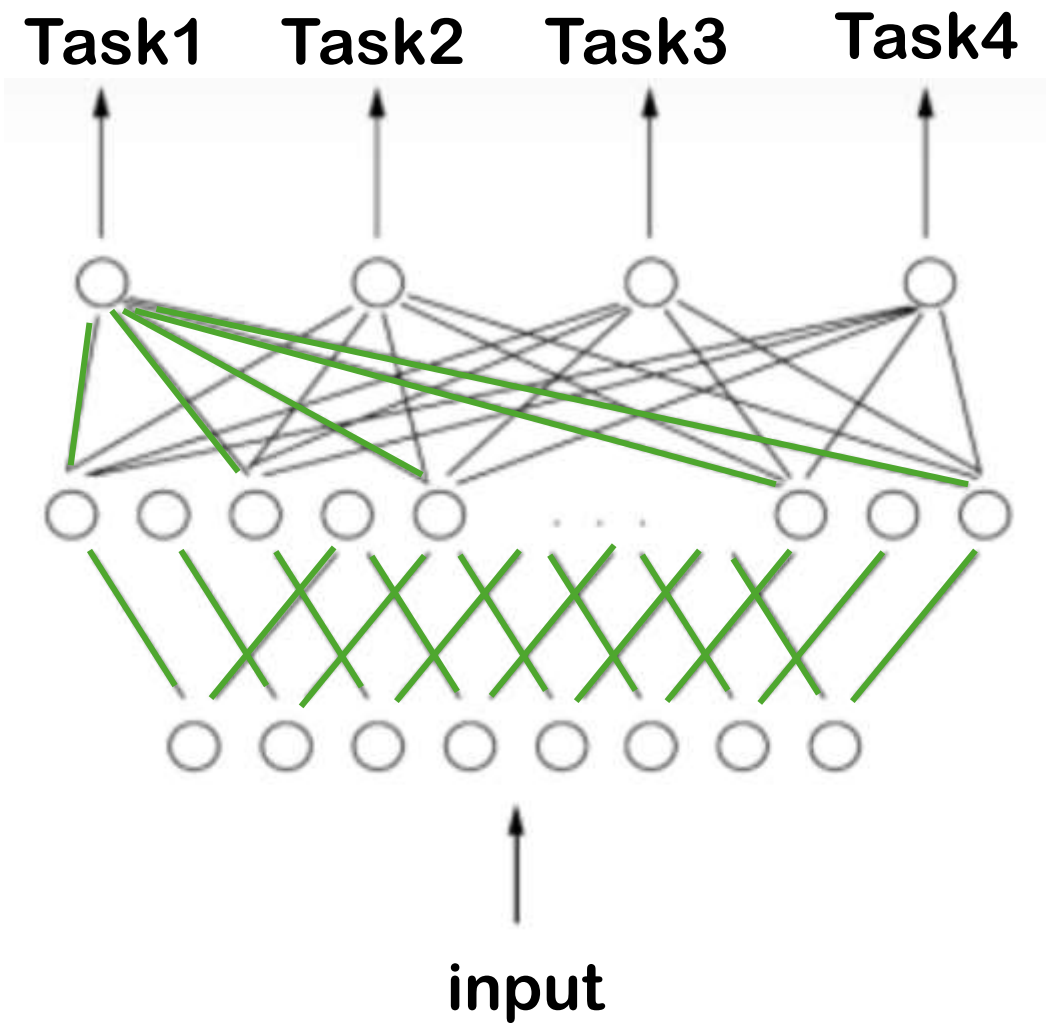
Task：针对不同群体的毒性检测

缺点：标签污染

- 例子：B群体不好
- Task[A]=无毒 Task[B]=有毒

Same Same, But Different: Conditional Multi-Task Learning for Demographic-Specific Toxicity Detection, WWW2023

条件多任务学习



	传统多任务学习		条件多任务学习	
文本数据	男生	女生	男生	女生
“我喜欢所有人”	无毒	无毒	无毒	无毒
“我喜欢男生”	无毒	无毒	无毒	∅
“我喜欢女生”	无毒	无毒	∅	无毒
“我讨厌所有人”	有毒	有毒	有毒	有毒
“我讨厌男生”	有毒	无毒	有毒	∅
“我讨厌女生”	无毒	有毒	∅	有毒

Same Same, But Different: Conditional Multi-Task Learning for Demographic-Specific Toxicity Detection, WWW2023

模型性能

		All		Men		Women	
		NT	T	NT	T	NT	T
Recall	Stacked STL	96.9	25.2	96.8	23.8	97.1	23.6
	CSMTL	97.2	23.6	98.8	13.3	99.7	4.6
	TradMTL	94.4	20.1	95.8	4.2	95.6	2.8
	CondMTL (Ours)	96.1	29.0	95.9	**28.7**	95.1	**31.2**
F1	Stacked STL	92.3	35.3	92.0	33.5	92.8	33.3
	CSMTL	92.3	33.8	92.2	22.2	92.8	8.7
	TradMTL	92.1	29.8	91.9	7.9	92.6	5.4
	CondMTL (Ours)	92.2	38.3	92.9	**37.9**	93.6	**39.5**
Precision	Stacked STL	88.2	58.6	87.6	56.9	88.9	56.4
	CSMTL	88.0	59.3	86.4	**67.0**	86.8	**69.1**
	TradMTL	87.5	54.4	85.3	47.7	86.6	46.7
	CondMTL (Ours)	88.6	56.1	88.2	55.9	89.7	53.7

Same Same, But Different: Conditional Multi-Task Learning for Demographic-Specific Toxicity Detection, WWW2023

毒性检测 – 减轻模型偏见

本篇总结

本文提出了条件多任务学习框架，增强了算法的公平性

提高了对于少数群体的有毒言论的检测性能

思路扩展

可以扩展到需要其他关注少数群体的问题

可以扩展到多维度（性别、民族、文化），甚至具体到个人

MuTox: Universal Multilingual Audio-based Toxicity Dataset and Zero-shot Detector, ACL2024

03. 结论

- 未来工作
- 国内研究团队

未来工作

1. 毒性检测模型需要实现的能力：
   - 实时性　（快速检测、实时训练）
   - 低成本　（配合其他任务实现，情绪分析）
   - 鲁棒性　（不需要针对于毒性检测）
   - 泛化能力（数据生成、数据预处理）
   - 跨语言能力（数据层面）

未来工作

2. 毒性检测模型（个人化 or 公众化）
   - 1. 不能用一个模型统领所有价值观
   - 2. 模型的针对人群至少应该群体化
   - 3. 群体划分问题，模型应该个人化

国内研究团队

朱福庆

中国科学院信息工程研究所副教授
在 iie.ac.cn 上验证电子邮件
多媒体信息检索　计算机视觉　深度学习　视觉与语言

- 通过跨领域知识转移实现多模态仇恨言论检测（2022）
- 用于仇恨言论检测的不确定性感知跨模态对齐（2024）
- 通过大型语言模型集成开放领域知识以实现多模态假新闻检测（2024）

国内研究团队

大连理工大学

软件与理论研究所

信息检索研究室 DUTIR

黄德根　李建明　李丽双　林鸿飞　林晓惠　罗 凌　孙媛媛　王 健　徐 博　徐喜荣

杨 亮　杨志豪　张绍武　周惠巍

21：　单语和多语恶意评论检测的混合模型

样本分布不均衡下的多语言恶意文本检测方法研究

Song gui zhe?

通过反事实因果效应消除恶意语言检测的偏见

使用深度学习检测社交媒体论坛中的自杀意念

通过神经多任务学习联合检测讽刺和幽默

Lin Boss、Xu(AI4sci+NLP)　Yang(NLP)

国内研究团队

宋阳秋

香港科技大学

cse.ust.hk 的电子邮件经过验证 -主页

人工智能　数据挖掘　自然语言处理　知识图谱　常识推理

- 探索大型预训练语言模型中的有毒内容（2021）
- 大型语言模型中的隐私：攻击、防御和未来方向（2023）
- ChatGPT 上的多步越狱隐私攻击（2023）
- 使用合成数据在大型语言模型上进行联合领域特定知识转移（2024）
- 生成大型语言模型的后门删除（2024）

国内研究团队

黄民烈

清华大学

tsinghua.edu.cn 的电子邮件经过验证 -主页

对话系统　自然语言生成　文本生成　情绪分析　自然语言处理

- 汉语大型语言模型安全性评估（2023）
- 揭示大型语言模型中的隐性毒性（2023）
- 安全忘却：一种出人意料的有效且可推广的防御越狱攻击的解决方案（2024）
- SafetyBench：评估大型语言模型的安全性（2024）
- 语言模型学会通过 RLHF 误导人类（2024）