计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2021.20200727
58(2):253-263,2021

基于社会新闻数据集的伦理行为判别方法

古天龙[1] 冯 旋[1] 李 龙[1,2] 包旭光[1] 李云辉[1]
[1](广西可信软件重点实验室(桂林电子科技大学) 广西桂林 541004)
[2](暨南大学信息科学技术学院/网络空间安全学院 广州 510632)
(gu@guet.edu.cn)

Ethical Behavior Discrimination Based on Social News Dataset

Gu Tianlong[1], Feng Xuan[1], Li Long[1,2], Bao Xuguang[1], and Li Yunhui[1]
[1](*Guangxi Key Laboratory of Trusted Software* (*Guilin University of Electronic Technology*), *Guilin*, *Guangxi* 541004)
[2](*College of Information Science and Technology/College of Cyber Security*, *Jinan University*, *Guangzhou* 510632)

Abstract With the broader applications of artificial intelligence (AI), their ethical and moral issues have attracted more and more concerns. How to develop an AI system that complies with human values and ethical norms from the perspective of technology realization, namely, ethical aligned AI design, is one of the important issues that need to be solved urgently. The ethical and moral discrimination based on machine learning is a beneficial exploration in this aspect. Social news data has rich content and knowledge of ethics and morality, which provides the possibility for the training data development of machine learning. Because of this, this paper constructs a social news dataset with ethics and morality of human behavior, which is attached to law and code of conduct dataset for machine learning training and testing. The ethical behavior discrimination model ERNIE-CNN based on enhanced language representation of information entities (ERNIE) and convolutional neural network (CNN), is developed to extract ethical discriminations about behavior by calculating semantic similarity based on the vector representation of words. The experimental results show that the proposed model has better performance than the baseline models.

Key words social news dataset; ethically aligned design; deep learning; ERNIE; CNN

摘 要 人工智能(artificial intelligence, AI)应用的伦理风险和挑战引起了人们的普遍关注,如何从技术实现角度开发出遵守人类价值观和伦理规范的AI系统,即,符合伦理的AI设计,是亟需解决的重要问题之一.基于机器学习的伦理与道德判别是此方面的有益探索.社会新闻数据具有丰富的伦理和道德的内容及知识,为机器学习的训练数据开发提供了可能.鉴于此,本文构建了具有人类行为伦理和道德属性的社会新闻数据集,附之以法律与行为规范数据集,用以机器学习的训练和测试;建立了基于使用信息实体的增强语言表示(enhanced language representation of information entities, ERNIE)和卷积

收稿日期:2020-09-15;**修回日期**:2020-10-30
基金项目:国家自然科学基金项目(U1711263,U1811264,61966009,61961007,61862016,62006057);广西自然科学基金项目(2019GXNSFBA245049,2019GXNSFBA245059,2018GXNSFDA281045);广西科技基地与人才专项(AD19245011)
This work was supported by the National Natural Science Foundation of China (U1711263, U1811264, 61966009, 61961007, 61862016, 62006057), the Natural Science Foundation of Guangxi Province (2019GXNSFBA245049, 2019GXNSFBA245059, 2018GXNSFDA281045), and the Science and Technology Base and Talents Program of Guangxi Province (AD19245011).
通信作者:李龙(lilong@guet.edu.cn)

神经网络(convolutional neural network, CNN)的伦理行为判别模型 ERNIE-CNN,通过词的向量表示计算语义相似度来提取关于行为的伦理判断.实验结果表明,提出的模型具有比基准模型更好的性能,验证了方法和模型的有效性.

关键词 社会新闻数据集;符合伦理的设计;深度学习;ERNIE;CNN

中图法分类号 TP391

随着深度学习技术的快速发展和物联网的不断更新,人工智能(AI)已经广泛应用到无人驾驶、智能家居、医疗护理等诸多领域,为生活带来巨大便利.但与此同时,人机交互方式的简化使得人们对技术的依赖程度越来越深,对其控制却逐渐减弱,技术的不完善和数据的不恰当使用引发的伦理问题不断出现[1-2].为了使技术更好地为人类服务,AI 系统需要具备伦理行为判别能力,理解特定社会的伦理道德规范,理解行为背后的社会、文化和伦理含义,才能在现实世界中进行大规模部署[3].

AI 系统的开发应当始终秉承以人为中心、造福全人类的发展理念,为其添加伦理行为判别能力成为日益紧迫的问题[4-6].通过嵌入伦理道德规范能够使 AI 系统具备伦理行为的判别及执行能力,从而遵守法律和社会行为规范,是 AI 具备道德能力的体现.但由于伦理概念的复杂性及当前技术发展的局限性,伦理行为判别方面的技术研究面临着诸多障碍.

早期针对伦理行为判别能力的研究工作主要包括基于规则的方法[7]和基于案例的方法[8],而这些方法普遍存在知识信息难以规则化、不具备自动获取特征表示的能力,导致模型的泛化能力并不理想.近年来,基于强化学习的方法[9]从试错或观察中学习人类的行为偏好,取得了很好的效果.虽然基于强化学习的方法行之有效,但是仍然具有以下不足之处:1)对环境变化的适应能力弱,只能适应特定领域而难以学习演示示例之外的知识;2)学习效率低下,示教者难以提供高质量的演示,而且提供演示需要耗费大量时间;3)模型训练严重依赖数据,往往需要海量的高质量数据,而且有反馈信号稀疏等问题,难以从原始数据中提取有用特征.

近期的研究表明,基于大规模无监督语料的预训练语言模型编码了行为规范等常识知识,包含丰富的语义信息,同时在高质量标注数据集上进行微调(fine-tuning)可以给目标任务带来巨大的效果提升[10-14].其具体优势有:1)近乎无限量的优质数据;2)一次学习多次复用;3)学习到的语义表征可在多个任务中进行快速迁移.

鉴于最近预训练语言模型的强大语义表示能力,针对缺乏高质量标注数据集的伦理行为判别任务,本文基于丰富的新闻语料构建了涵盖伦理道德和人类行为的社会新闻数据集,同时基于使用信息实体的增强语言表示(enhanced language representation with informative entities, ERNIE)[14-16]和卷积神经网络(convolutional neural network, CNN)[17]提出了伦理行为判别模型 ERNIE-CNN.通过伦理行为判别实验和零样本迁移实验,证明了本文所提方法和模型的有效性,同时为 AI 伦理的未来研究提供必要的支持.

本文的主要贡献有 3 个方面:

1) 基于丰富的新闻语料构建了社会新闻数据集,为了验证伦理判别方法的有效性,同时构建了法律与行为规范数据集.

2) 基于使用信息实体的增强语言表示(ERNIE)和卷积神经网络(CNN)提出 ERNIE-CNN 模型.

3) 通过零样本迁移实验验证了协变量偏移下的领域适应能力,验证伦理行为判别方法的有效性.

1 相关工作

为 AI 系统增添伦理行为判别能力,有助于避免伦理问题,提高人类对技术的接受程度,从而促进 AI 技术的研究与推广.早期的相关研究工作大多选择将明确的伦理原则或规范嵌入 AI 系统,主要包括基于规则的方法和基于案例的方法.

基于规则的方法是根据伦理原则或规范为 AI 系统嵌入伦理行为判别的能力,实现了不同的伦理原则.Anderson 等人[18]基于罗尔斯义务论开发了伦理顾问系统 W.D.,该系统从案例数据中学习各项行为原则的权重,计算并选择具有最高效用值的行为.Anderson 等人[19]基于生物医学理论设计了伦理顾问系统 MedEthEx,使用归纳逻辑程序设计技术提取伦理信息,辅助医护人员确定伦理行为选择,用于解决医疗护理中涉及人机交互的伦理问题.Arkin[20]

将战争法和交战规则嵌入自主武器系统,用于约束系统的致命行为,判断其行为是否符合伦理道德.尽管上述研究工作实现了对单一伦理原则或规范的嵌入,但并未考虑不同规范间存在的规范冲突风险.总体而言,基于规则的方法具有较强的可解释性,能使人们更直观地了解模型的判别过程,而无法推理未编码到知识库中的条件和规则,难以避免规范冲突问题,因此其性能表现较为受限.

基于案例的方法是通过重用以往经验进行规范嵌入,通过类比以往的案例自动提取规范并求解问题的方法.Ashley 和 McLaren[21-22]使用基于案例的方法设计了 SIROCCO 程序,探索和分析案件遵循的伦理原则和具体事实之间的关系.Dehghani 等人[23]提出了计算模型 MoralDM.当面对新的伦理决策场景时,MoralDM 中的类比推理模块将新场景与数据库中先前已解决的多个场景进行比较,计算新案例和已解决方案之间的相似度并以此为基础进行类比推理.但随着案例数量的增加,MoralDM 穷举比较方法的计算复杂度将相应增加,导致实用性变差.Blass 等人[24]利用结构映射扩展了 MoralDM 模型,通过计算案例及候选规范之间的相似度缩小搜索空间,提高基于案例方法的效率.总体而言,基于案例的方法具有信息表达完整、求解方法简单等优点,已经在伦理行为判别研究中成功运用,但是其缺点也较为突出,即有限的规范无法适应不断变化的道德场景,而且必须解决案例的相似度度量、训练案例的选取等问题.

近年来,基于强化学习的方法通过人类专家示教、奖励的形式学习人类的行为偏好,同样能够为 AI 系统嵌入伦理行为判别能力.强化学习通常使用动态规划技术来解决问题,以试错的方式自主学习或从示教者提供的示例中学习,以达到奖励函数最大化并最终实现特定目标.Abel 等人[25]将强化学习形式化为部分可观察马尔可夫决策过程(POMDP),并针对 2 个伦理困境(蛋糕或死亡、火灾救援实验)验证了该方法的灵活性和稳定性.然而在有限步(Finite Horizon)内,POMDP 问题难以求解.Wu 等人[26]基于反向强化学习,利用专家行为数据,通过最大化奖励函数来平衡道德行为和效用值的追求.但由于每个人的偏好不同,专家的行为数据也可能存在分歧.Riedl 等人[27]认为一个可以阅读和理解故事的模型能够从故事所直接体现出的或隐含的社会文化知识中学习其所蕴含的社会行为规范,因此基于强化学习提出了从众包故事中学习人类行为偏好的方法.此方法的局限性在于众包故事的获取途径单一,只适用于特定任务.总体而言,基于强化学习的方法能够通过与环境交互、学习获得人类行为偏好,具备解决复杂问题的能力,但此类方法在很大程度上依赖于输入信号的质量,制约了该方法的性能表现及实际应用.

近期的研究表明,使用大规模无监督文本数据进行预训练的语言模型编码了文本中行为规范等常识概念[28-29].Ziegler 等人[30]使用预训练语言模型 GPT-2 成功学习到人类生成句子的偏好,验证了语言模型可以从文本数据中学到人类行为偏好.Frazier 等人[31]使用长期连载的儿童漫画构建数据集,通过训练语言模型识别文本内容是否符合社会规范.尽管上述工作验证了语言模型蕴含知识信息,但过分强调精心策划的场景,难以应对现实场景带来的挑战.

综合上述研究,本文旨在为伦理行为判别研究提供新的方法以及在语言模型上进行改进.鉴于社会新闻充分涵盖伦理道德且易于获取,构建了社会新闻数据集用于伦理行为判别研究.提出了伦理行为判别模型 ERNIE-CNN,从大量的社会新闻数据中学习行为偏好,通过词向量表示计算语义相似度来提取关于行为的伦理判断,解决了场景限制问题.经过训练的模型编码了知识信息,可以理解行为背后的社会、文化和伦理含义.

2 数据集构建

为了推动因缺乏高质量标注数据而受阻的伦理行为判别研究,本文选择充分涵盖伦理道德和人类行为的社会新闻标题为数据源,构建了社会新闻数据集(ETH-News),并将新闻文本中包含的行为分类为道德行为、不文明行为、违规行为和违法行为,用于伦理行为判别模型的训练.同时构建了法律与行为规范数据集(ETH-Norms),用于验证伦理行为判别方法的有效性.以上数据集的构建均包括数据采集、数据标注和数据集分析 3 个阶段.

2.1 数据采集

2.1.1 社会新闻数据集(ETH-News)

本文选择社会新闻标题作为伦理行为的主要数据源,原因主要有 3 点:1)社会新闻充分涵盖伦理道德和人类行为,同时具有易于获取的优点.2)社会新闻以较简明扼要的文字,向公众传达重要信息,每条新闻还有一个非常详实且简短的新闻摘要.与新闻

全文相比,新闻摘要内容丰富、简明扼要.3)社会新闻是涉及人民群众日常生活的社会事件、社会问题、社会风貌的报道,具有公开性、真实性、时效性、准确性和广泛性等特点.

本文针对中文语境中的伦理行为判别任务进行建模,采集的新闻文本主要爬取于新浪微博①,并使用 THUCTC 工具包②筛选出社会新闻.THUCTC 是由清华大学自然语言处理实验室推出的中文文本分类工具包,能够自动高效地实现用户自定义的文本分类任务.由于负面新闻比正面新闻更容易吸引大众的注意力,因此媒体登载了较多的负面新闻,导致自动采集的新闻文本存在样本不平衡问题.为了解决这一问题,本文进一步从中国文明网③的好人好事专栏爬取了全部的新闻标题,以此扩充正面新闻,因为好人好事是道德行为理想的数据来源.

本文期望通过具体的行为训练模型伦理行为判别的能力,因此对上述新闻进行了筛选,只保留了至少包含一个具体行为的文本,同时删除:1)不包含具体行为的文本;2)字段长度超过 52 个字符的文本;3)格式错误的文本.经过数据清理,使用剩余的 12 183 条新闻文本构建了所需社会新闻数据集.表 1 为社会新闻数据集的示例展示:

Table 1 Samples From ETH-News Dataset

表 1 社会新闻数据集示例

社会新闻标题	数据来源	类别	标签
市民见义勇为围追堵截持刀劫匪	社会新闻	Moral	0
大学食堂以豆制品冒充猪肉	社会新闻	Immoral	1
两名少年闹市抢劫,为达目的用铁棍敲受害人脑袋	社会新闻	Illegal	2

2.1.2 法律与行为规范数据集(ETH-Norms)

为了验证基于社会新闻数据集训练的模型具备伦理行为判别能力,本文同时构建了法律与行为规范数据集来验证模型识别法律与行为规范的能力.本文选择将《中华人民共和国刑法》与各省市《文明行为条例》纳入数据集,将各项条款拆分为简单句,文本长度同样控制在 52 字以内.表 2 为法律与行为规范数据集中拥有不同标签的示例展示.

Table 2 Samples From ETH-Norms Dataset

表 2 法律与行为规范数据集示例

法律与行为规范	数据来源	类别	标签
行人通过路口或者横过道路遇机动车礼让时快速通过	文明条例	Moral	0
行人不按照交通信号通行,乱穿马路,翻越交通护栏	文明条例	Immoral	1
对正在进行的暴力犯罪,采取防卫行为,造成不法侵害人伤亡	法律条文	Moral	0
故意杀人、故意伤害致人重伤或者死亡	法律条文	Illegal	2

2.2 数据标注

数据集的标注工作由实验室九名硕士研究生共同完成,男女比例为 5∶4.九名硕士研究生平均分为 3 组,每组 3 人,每条新闻文本由组内 2 人进行标注,另一人为仲裁.当 2 人标注结果相同时则完成标注(占总数的 93%),如有分歧,由仲裁者进行仲裁(占总数的 6.9%).在 3 人都难以标注的情况下丢弃样本(占总数的 0.1%),以此在最大程度上保证标注的一致性和准确性.

功利主义是一种主张最大化所有人的总体幸福感的理论[4].为了衡量新闻中包含的行为是否合乎伦理道德,本文选择功利主义为道德评判标准,计算公式为[32]

$$M=\sum_{i=1}^{N}(\boldsymbol{W}_i\times P_i), \tag{1}$$

其中,N 是利益相关者的数量,$\boldsymbol{W}_i$ 是每个利益相关者的权重,P_i 用于衡量每个利益相关者的幸福度,M 是某一行为的效用值.

每个标注人员根据功利主义将文本标注为 0(道德行为),1(不文明行为),2(违规行为),3(违法

① https://weibo.com/

② http://thuctc.thunlp.org/

③ http://www.wenming.cn/

行为),4(无关行为),同时删除所有不包括具体行为的新闻文本.

2.3 数据集分析

本文标注了 16 000 条新闻文本,修改了 448 条法律条文与 20 个省市的文明行为条例,经过数据清理等步骤,进一步删除了属于无效类别的所有文本.所构建的社会新闻数据集包含 3 496 条道德行为(Moral)、1 777 条不文明行为(Uncivilized)、994 条违规行为(Violative)和 5 916 条违法行为(Illegal).由于不文明行为和违规行为标签下的数据量少且样本表现形式不易区分,参考《文明行为条例》将其进行合并为不道德行为(Immoral).所构建的法律与行为规范数据集包含 283 条道德行为、149 条不道德行为和 709 条违法行为.

数据集的详细统计信息如表 3 所示,文本长度分布如图 1 所示,大部分文本长度小于 30 个字符,其中社会新闻数据集平均文本长度为 17.1,法律与行为规范数据集平均文本长度为 20.

Table 3 Statistics of Datasets

表 3 数据集的统计信息

标签	ETH-News	ETH-Norms
Moral	3 496	283
Immoral	2 771	149
Illegal	5 916	709
总计	12 183	1 141

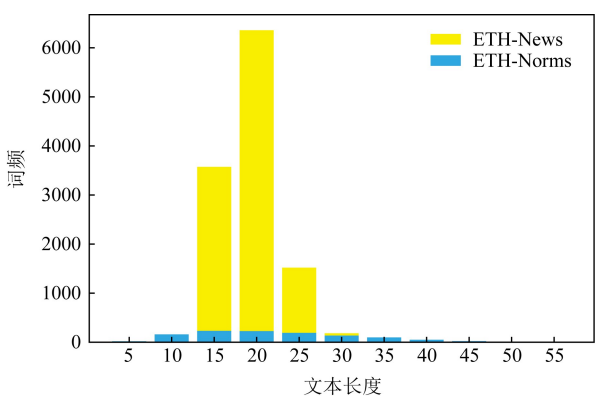

Fig. 1 Length distribution of text

图 1 文本长度分布

3 ERNIE-CNN 伦理行为判别模型

为了使模型具备伦理行为判别能力,理解行为背后的社会、文化和伦理含义,本文提出了 ERNIE-CNN 伦理行为判别模型,其整体架构如图 2 所示,主要由 4 部分组成,分别为词嵌入层、文本卷积层、池化层和输出层.

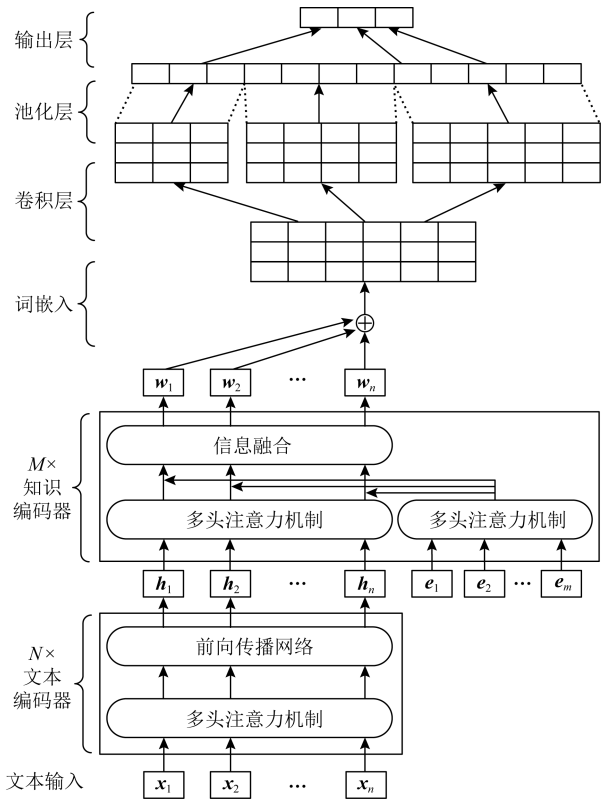

Fig. 2 ERNIE-CNN architecture

图 2 ERNIE-CNN 架构

3.1 词嵌入层

词嵌入层主要实现的功能是基于语言模型 ERNIE[14] 与输入文本进行交互,生成文本的词向量表示.ERNIE 针对 BERT[13] 在处理中文文本时难以获得语义完整表示的缺点,为了抽取和编码知识信息,将知识模型中的实体表征整合到语义模型的底层中,结合大规模无监督语料库和知识图谱进行预训练.

词嵌入层由 2 个模块组成:1)文本编码器(T-Encoder),负责从输入的文本中捕获词汇和语义信息;2)知识编码器(K-Encoder),负责将知识图谱中的知识信息整合到输出的词向量中.

文本编码器是包含多头注意力机制和前馈神经网络的多层双向 transformer[33] 编码单元,其架构如图 3 所示.

给定一条社会新闻,令 $\boldsymbol{x}_i \in \mathbb{R}^{ik}$ 为句子中第 i 个单词所对应的 k 维词向量,通过式(2)计算每个字符的词汇和语义特征:

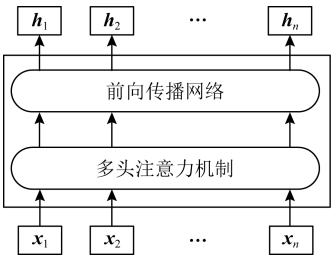

Fig. 3 T-Encoder architecture

图 3 文本编码器架构

$\{\boldsymbol{h}_1,\boldsymbol{h}_2,\cdots,\boldsymbol{h}_n\}=\text{T-Encoder}(\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\})$, (2)

其中,$\{\boldsymbol{h}_1,\boldsymbol{h}_2,\cdots,\boldsymbol{h}_n\}$代表具有语义特征的词嵌入输出;$n$ 为文本长度.

多头注意力机制(multi-head attention)[33]是文本编码器中的核心组成单元之一.注意力机制具有快速提取关键信息的重要特征,广泛应用于自然语言处理任务.注意力机制可以描述为一个查询(query)到一系列键-值对(key-value)的映射.注意力机制计算方法如下[33]:

$$Att(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V}, \tag{3}$$

其中,分别用向量 $\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}$ 表示查询和键-值对.首先,将 $\boldsymbol{Q}$ 和 $\boldsymbol{K}$ 进行相似度计算(点积)得到权重,为了防止点乘结果数值过大,使用向量 $\boldsymbol{K}$ 的维度 d_k 进行缩放;其次,使用 softmax 函数对权重进行归一化得到概率分布;最后,将权重与相应的键值 $\boldsymbol{V}$ 进行加权求和得到目标的 Attention.在自然语言处理任务中,$\boldsymbol{K}$ 通常与 $\boldsymbol{V}$ 取值相同,即 $\boldsymbol{K}=\boldsymbol{V}$.

自注意力机制是注意力机制的改进,为了捕获句子的内部相关性,减少对外部信息的依赖.在自注意力机制中,$\boldsymbol{Q}=\boldsymbol{K}=\boldsymbol{V}$.

多头注意力机制利用多个查询,并行地从输入信息中选取多组信息,可以提取多重语义的含义.多头注意力机制将数据投影到 h(注意力机制头数)个子空间中,考虑了多个子空间中向量的相似度.其中单头注意力 H_i 的计算公式如式(4)所示[33]:

$$H_i=Att(\boldsymbol{Q}\boldsymbol{W}_i^{Q},\boldsymbol{K}\boldsymbol{W}_i^{K},\boldsymbol{V}\boldsymbol{W}_i^{V}), \tag{4}$$

其中,$\boldsymbol{W}_i^{Q},\boldsymbol{W}_i^{K},\boldsymbol{W}_i^{V}$ 为 $\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}$ 的权重矩阵.

多头注意力机制将所有空间中的注意力向量进行拼接,计算公式如式(5)所示[33]:

$$MH\text{-}Att(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=Concat(H_1,H_2,\cdots,H_h)\boldsymbol{W}^{0}, \tag{5}$$

其中,$\boldsymbol{W}^{0}$ 是附加权重矩阵,作用是将拼接后的矩阵维度压缩成固定的文本长度大小.

知识编码器可以编码字符和实体,也能融合异构特征,作用是将知识信息注入语义表征,其结构如图 4 所示:

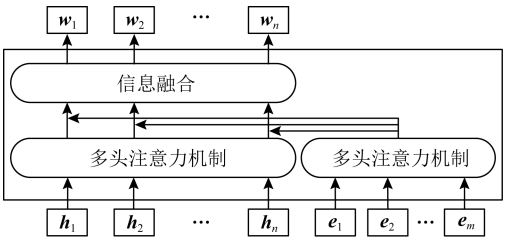

Fig. 4 K-Encoder architecture

图 4 知识编码器架构

在知识编码器中,将字符嵌入$\{\boldsymbol{h}_1,\boldsymbol{h}_2,\cdots,\boldsymbol{h}_n\}$和预训练得到的实体嵌入$\{\boldsymbol{e}_1,\boldsymbol{e}_2,\cdots,\boldsymbol{e}_m\}$(其中 m 是实体对齐序列长度)进行异构信息融合并通过式(6)计算得到最终的输出词嵌入$\{\boldsymbol{w}_1,\boldsymbol{w}_2,\cdots,\boldsymbol{w}_n\}$.

$$\{\boldsymbol{w}_1,\boldsymbol{w}_2,\cdots,\boldsymbol{w}_n\}=\text{K-Encoder}(\{\boldsymbol{h}_1,\boldsymbol{h}_2,\cdots,\boldsymbol{h}_n\},\{\boldsymbol{e}_1,\boldsymbol{e}_2,\cdots,\boldsymbol{e}_m\}). \tag{6}$$

3.2 文本卷积层

在获得词嵌入输出$\{\boldsymbol{w}_1,\boldsymbol{w}_2,\cdots,\boldsymbol{w}_n\}$后,通过文本卷积操作提取句子的局部区域特征,能够自动地对 N-gram 特征进行组合和筛选,获得不同抽象层次的语义信息.文本卷积层结构如图 5 所示:

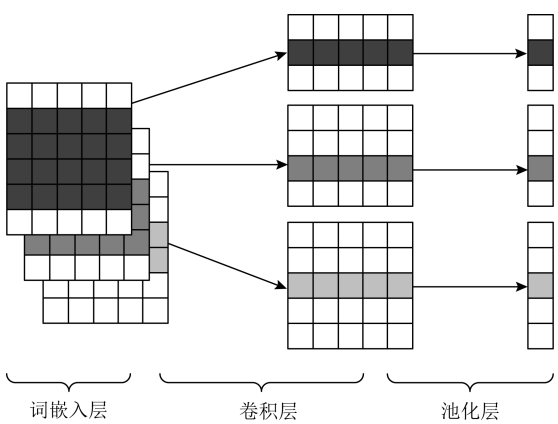

Fig. 5 Example of text convolution process

图 5 文本卷积过程示例

对于输入的每一个句子 $\boldsymbol{s}$,将 $\boldsymbol{s}$ 中字符的词向量$\{\boldsymbol{w}_1,\boldsymbol{w}_2,\cdots,\boldsymbol{w}_n\}$进行连接操作,如式(7)所示.

$$\boldsymbol{s}=\boldsymbol{w}_1\oplus\boldsymbol{w}_2\oplus\cdots\oplus\boldsymbol{w}_n, \tag{7}$$

其中,$\oplus$表示词向量间的连接操作.$\boldsymbol{s}$ 是拼接得到的 $n\times k$ 维矩阵,其中 n 为一个句子中的单词数,k 是每个单词对应的词向量维度.

卷积层使用卷积核 $W\in\mathbb{R}^{jk}$ 与滑动窗口 $\boldsymbol{s}_{i:i+j-1}$ 对输入的 $n\times k$ 维矩阵进行卷积操作,产生特征 $\boldsymbol{c}_i$:

$$\boldsymbol{c}_i = f(\boldsymbol{W} \cdot \boldsymbol{s}_{i:i+j-1} + b), \tag{8}$$

其中,j 表示窗口中的单词数,$\boldsymbol{s}_{i:i+j-1}$ 代表由输入矩阵的第 i 行到第 $i+j-1$ 行拼接而成的大小为 $j \times k$ 维的窗口,$\boldsymbol{W}$ 为卷积核对应的 $j \times k$ 维的权重矩阵,b 为偏置参数,$f(\cdot)$ 为非线性函数 tanh.首先,$\boldsymbol{W}$ 和 $\boldsymbol{s}_{i:i+j-1}$ 进行点积运算;其次,卷积核对 j 个字符按照步长 1 滑动提取文本局部特征 $\boldsymbol{c}_i$;最后,将其拼接得到特征映射 $\boldsymbol{c} = (\boldsymbol{c}_1, \boldsymbol{c}_2, \cdots, \boldsymbol{c}_{n-j+1})$.

3.3 池化层

经卷积操作后,使用时序最大池化操作(1-Max Pooling)从每个滑动窗口产生的特征向量中筛选出最重要的特征,并将这些特征进行拼接,构成向量表示.

3.4 输出层

得到文本句子的向量表示之后,接入全连接层输出每个类别的概率,并使用 softmax 激活函数进行归一化处理,得到最终分类结果.

4 实验设计与结果分析

本文在社会新闻数据集上训练模型,对社会新闻中包含的行为进行伦理行为判别实验,其中二分类(行为被划分为道德、不道德 2 类)实验可以体现模型辨别对错的能力,三分类(行为被划分为道德、不道德、违法 3 类)具有更细粒度的分类能力,体现模型识别行为伦理程度的能力,同时使用训练好的模型在法律与行为规范数据集上验证该方法的有效性.

4.1 度量标准

伦理行为判别属于多分类问题,为了计算分类模型在不同类别上的总体精确率、召回率和 $F1$ 值,本文选用宏精确率($macro\text{-}P$)、宏召回率($macro\text{-}R$)和宏 $F1$ 值($macro\text{-}F1$)作为模型预测评价指标,如式(9)~(11)所示:

$$macro\text{-}P = \frac{1}{K}\sum_{i=1}^{K} P_i, \tag{9}$$

$$macro\text{-}R = \frac{1}{K}\sum_{i=1}^{K} R_i, \tag{10}$$

$$macro\text{-}F1 = \frac{1}{K}\sum_{i=1}^{K} \frac{2 \times P_{\text{macro}} \times R_{\text{macro}}}{P_{\text{macro}} + R_{\text{macro}}}. \tag{11}$$

其中,K 为类别数;P_i 为精确率;R_i 为召回率.

4.2 数据与设置

本文共设置 2 个实验,在伦理行为判别实验中(实验 1),社会新闻数据集的划分情况如表 4 所示(在二分类任务中,将不道德行为与违法行为划分为一类),在零样本迁移实验中(实验 2),将法律与行为规范作为测试集,数据集的划分情况如表 5 所示.

Table 4 Datasets Used in Experiment 1

表 4 实验 1 使用的数据集

标签	训练集	验证集	测试集
Moral	2 496	500	500
Immoral	1 771	500	500
Illegal	3 916	1 000	1 000
总计	8 183	2 000	2 000

Table 5 Datasets Used in Experiment 2

表 5 实验 2 使用的数据集

标签	训练集	验证集	测试集
Moral	2 996	500	283
Immoral	7 187	1 500	858
总计	10 183	2 000	1 141

本文采用基于字符的方法对数据进行预处理.实验中,基于特征的语言模型使用搜狗预先训练好的 300 维中文词向量① 初始化单词嵌入,基于微调的语言模型使用谷歌和百度发布的预训练模型.短填长切,将每句话长度处理为 30,设置 batch_size 为 64.采用 Adam 梯度下降更新网络参数,使用交叉熵(cross entropy)作为代价函数,计算模型预测结果与训练实例之间的差异性.为了缓解训练中可能出现的过拟合问题,在模型的全连接层使用了随机失活(Dropout)和提前停止技术.实验代码基于 Pytorch1.5.1 实现.

4.2.1 伦理行为判别实验

为了验证社会新闻数据集在训练模型进行伦理行为判别方面的有效性,将新闻文本中的行为分为道德行为、不道德行为和违法行为,并使用 8 种不同的基准模型来构建二元、三元分类器.

1) TextCNN[17].本文选择卷积核的大小为(2,3,4),每个尺寸的卷积核个数为 100,dropout 设置为 0.5,学习率设置为 1×10^{-3}.

2) TextRNN[34](Bi-LSTM).使用双向长短期记忆网络捕捉长距离语义关系,能够更好的表达上下文信息.TextRCNN[35] 在 Bi-LSTM 的基础上加入一层最大池化层来捕捉重要的特征信息.本文使用 2 层双向 LSTM,每层包含 256 个神经元,dropout 设置为 0.5,学习率设置为 1×10^{-4}.

① https://github.com/Embedding/Chinese-Word-Vectors

3) RNN-Att[36].在 Bi-LSTM 的基础上加入注意力机制,能够直观的解释各个句子和词对分类类别的重要性.本文使用 2 层双向 LSTM,第 1 层包含 128 个神经元,第 2 层包含 64 个神经元,dropout 设置为 0.5,学习率设置为 1×10^{-3}.

4) DPCNN[37].引入了残差结构,增加了多尺度信息,并且增加了用于文本卷积神经网络的网络深度,以提取文本中远程关系特征.本文选择卷积核的大小为 3,卷积核个数为 250,dropout 设置为 0.5,学习率设置为 1×10^{-5}.

5) BERT[13].采用多层双向 Transformer 结构以及掩码语言模型来捕捉一个词在上下文语境中的词向量表达,极大程度提升了词向量的表征能力.本文使用谷歌发布的预训练模型 BERT_Chinese① 进行微调,微调的学习率设置为 5×10^{-5}.

6) BERT-CNN[38].将 BERT 的输出作为词嵌入层,再由卷积神经网络经过卷积后提取句子级别特征,经过池化层保留重要特征,微调的学习率设置为 5×10^{-5}.

7) BERT-RCNN.基于 BERT-CNN 与 TextRCNN 的设计思路,将 BERT 的输出作为词嵌入层,加入 Bi-LSTM 捕捉长距离语义关系,最后加入一层最大池化层来捕捉层次语义信息,微调的学习率设置为 5×10^{-5}.

8) ERNIE-CNN.本文使用百度发布的预训练模型 ERNIE_Chinese② 进行微调,微调的学习率设置为 5×10^{-1}.

4.2.2 零样本迁移实验

为了验证基于社会新闻数据集伦理行为判别方法的有效性,本文通过零样本迁移实验进行验证,测试模型在训练集和测试集上分布不同的协变量偏移下的领域适应能力.实验 2 使用在实验 1 中性能表现更出色的基于微调的语言模型,训练集为法律与行为规范数据集.实验 2 中其他参数设置与实验 1 相同.

4.3 实验结果分析

伦理行为判别实验结果如表 6 所示.零样本迁移实验结果如表 7 所示.本文着重分析更细粒度的三分类实验结果,其更能体现模型伦理行为判别的能力.

Table 6 Ethical Behavior Discrimination Experiment Results

表 6 伦理行为判别实验结果

模型划分	模型	二分类			三分类		
		宏精确率	宏召回率	宏 $F1$ 值	宏精确率	宏召回率	宏 $F1$ 值
基于特征	TextCNN	**0.907**	**0.889**	**0.898**	**0.764**	**0.744**	**0.750**
	TextRCNN	0.896	0.883	0.889	0.740	0.720	0.726
	TextRNN-ATT	0.899	0.885	0.891	0.739	0.718	0.722
	DPCNN	0.894	0.884	0.889	0.733	0.729	0.729
基于微调	BERT	0.951	0.944	0.948	0.807	0.766	0.773
	BERT-CNN	0.934	0.947	0.940	0.819	0.781	0.785
	BERT-RCNN	0.956	0.952	0.954	0.815	0.763	0.769
	ERNIE-CNN	**0.957**	**0.966**	**0.961**	**0.829**	**0.808**	**0.813**

注:黑体数字表示最佳性能.

Table 7 Zero-shot Transfer Experiment Results

表 7 零样本迁移实验结果

模型	二分类			三分类		
	宏精确率	宏召回率	宏 $F1$ 值	宏精确率	宏召回率	宏 $F1$ 值
BERT	0.839	0.875	0.854	0.695	0.735	0.689
BERT-CNN	0.858	0.854	0.856	0.708	0.758	0.711
BERT-RCNN	0.850	0.879	0.863	0.696	0.748	0.705
ERNIE-CNN	**0.890**	**0.929**	**0.906**	**0.767**	**0.803**	**0.781**

注:黑体数字表示最佳性能.

4.3.1 伦理行为判别实验结果分析

通过对比不同基于特征的语言模型(TextCNN,TextRCNN,TextRNN-ATT 和 DPCNN)与基于微调的语言模型(BERT,BERT-CNN,BERT-RCNN 和 ERNIE-CNN)下得出的实验结果可以看出,在语义表示能力方面,基于微调的语言模型优于基于特征的语言模型,表明基于 transformer 结构的模型效果明显优于非基于 transformer 结构的模型.

① https://github.com/google-research/bert

② https://github.com/PaddlePaddle/ERNIE

二分类任务.通过二分类任务的实验结果可以看出,本文提出的 ERINE-CNN 模型在 *macro-P*,*macro-R* 和 *macro-F*1 指标上均获得了最佳性能,分别达到了 0.957,0.966 和 0.961,效果最差的基准模型为 DPCNN,在 *macro-P*,*macro-R* 和 *macro-F*1 指标上分别为 0.894,0.884 和 0.889.即使效果最差的模型也取得了令人满意的实验结果,表明基于社会新闻数据集训练的伦理行为判别模型可以对社会新闻中包含的行为准确地分类.

三分类任务.通过对比不同基于特征的语言模型下得出的实验结果可以看出,尽管 TextCNN 模型相较于其他模型结构更为简单,但在 3 个衡量指标上均获得了最佳效果,分别达到了 0.764,0.744 和 0.750.由于 DPCNN 是为捕捉长距离语义关系而设计的模型,因此在本文的短文本任务中效果较差.通过对比 TextCNN 与基于 TextRNN 改进的 TextRCNN 和 TextRNN-ATT 的实验结果可以发现,模型的堆叠并没有为 TextRNN 带来效果的提升,反而增加了模型的复杂程度.本实验中针对基于特征的语言模型所进行的对比表明,对于短文本伦理行为判别任务,由于句子不具有复杂的结构,通过滑动卷积操作捕捉局部特征就可以识别全局语句结构,获取句子中最重要的语义信息,在任务中获得较好的结果.

通过对比基于微调的语言模型的实验结果可以发现,本文提出的 ERINE-CNN 模型在 *macro-P*,*macro-R* 和 *macro-F*1 均获得了最佳效果,分别达到了 0.829,0.808 和 0.813.通过对比 BERT,BERT-CNN 和 BERT-RCNN 的实验结果发现,将 BERT 的输出作为词嵌入层,加入 CNN 和 RCNN 模型之后在各项指标上均有所提升.BERT-CNN 加入 CNN 模型后在 *macro-P*,*macro-R* 和 *macro-F*1 指标上分别提升了 1.2,1.5 和 1.2 个百分点,高于 BERT-RCNN 加入 RCNN 模型带来的效果提升,同时再次验证加入滑动卷积操作捕捉局部特征的有效性.通过对比 BERT-CNN 和本文提出的 ERNIE-CNN 模型下的实验结果,可以发现 ERINE-CNN 模型在 *macro-P*,*macro-R* 和 *macro-F*1 指标上均获得了更好的性能,与 BERT-CNN 相比分别提升了 1,2.7 和 2.8 个百分点,表明在处理中文任务时,ERNIE 的特征抽取能力比 BERT 更强,同时也证明了本文提出的模型在伦理行为判别任务中的有效性.

4.3.2 零样本迁移实验结果分析

通过二分类任务的实验结果可以看出,本文提出的 ERINE-CNN 模型在 *macro-P*,*macro-R* 和 *macro-F*1 均获得了最佳效果,分别达到了 0.890,0.929 和 0.906,效果最差的基准模型 BERT 在 *macro-P*,*macro-R* 和 *macro-F*1 指标上分别为 0.839,0.875 和 0.854;对于三分类任务,ERNIE-CNN 模型同样达到了最佳效果,在各项指标上分别达到了 0.767,0.803 和 0.781,效果最差的基准模型在各项指标上分别达到了 0.695,0.735 和 0.689.实验结果表明使用社会新闻数据集训练好的模型在零样本迁移实验上同样可以取得令人满意的实验结果,在资源有限的情况下,验证了社会新闻数据集蕴含丰富的伦理道德和规范知识,可以用于伦理行为判别研究.

5 总 结

本文针对包含具体行为的社会新闻做伦理行为判别研究,受预训练语言模型的启发,提出了基于社会新闻数据集的伦理行为判别方法.具体地,由于缺乏高质量标注数据,基于社会新闻和社会规范分别构建了社会新闻数据集和法律与行为规范数据集.使用预训练语言模型 ERNIE 捕获新闻文本的多维语义特征获得词向量,通过 CNN 模型自动对 *N*-gram 特征进行组合和筛选,获得不同抽象层次的语义信息,从而提升模型的识别能力.在数据集上的实验结果证明了以上方法和模型的有效性.

具备伦理行为判别能力的模型可以为机器人、无人驾驶汽车等自主机器提供先验知识,减少不道德行为的发生,有助于避免伦理问题.本文的研究工作为伦理行为判别开辟了新思路,是一次有益的尝试.

然而,本文工作也存在一些问题:1)自动采集的新闻文本在不同标签下存在数据量少和数据不均衡问题,而本文仅对不易区分的数据进行合并;2)仅针对短文本任务进行优化,未考虑模型对长文本的识别能力;3)社会新闻是被精心编辑过的规范数据,而社交媒体领域有大量含有噪声的非规范数据.在接下来的研究工作中,我们将补充更多的数据,扩展模型更细粒度的伦理行为辨别能力;优化模型对于长文本的识别能力;探索从新闻领域到社交媒体领域的迁移学习任务.

参 考 文 献

[1] Wallach W, Asaro P. Machine ethics and robot ethics [M]. London: Routledge, 2017

[2] Allen C, Wallach W, Smit I. Why machine ethics? [J]. IEEE Intelligent Systems, 2006, 21(4): 12-17

[3] Lourie N, Bras R L, Choi Y. Scruples: A corpus of community ethical judgments on 32,000 real-life anecdotes [J]. arXiv preprint, arXiv:2008.09094, 2020

[4] Wallach W, Allen C. Moral machines: Teaching robots right from wrong [M]. Oxford: Oxford University Press, 2008

[5] Dennis L A, Fisher M, Slavkovik M, et al. Formal verification of ethical choices in autonomous systems [J]. Robotics and Autonomous Systems, 2016, 77(33): 1-14

[6] Arnold T, Kasenberg D, Scheutz M, et al. Value alignment or misalignment - What will keep systems accountable? [C] //Proc of the 31st AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2017: 1-8

[7] Weiss S M, Indurkhya N. Rule-based machine learning methods for functional prediction [J]. Journal of Artificial Intelligence Research, 1995, 3(1): 383-403

[8] Aamodt A, Plaza E. Case-based reasoning: Foundational issues, methodological variations, and system approaches [J]. AI Communications, 1994, 7(1): 39-59

[9] Sutton R S, Barto A G. Reinforcement learning: An introduction [J]. IEEE Transactions on Neural Networks, 1998, 9(5): 1054-1054

[10] Peters M E, Neumann M, Iyyer M, et al. Deep contextualized word representations [J]. arXiv preprint, arXiv: 1802.05365, 2018

[11] Howard J, Ruder S. Universal language model fine-tuning for text classification [J]. arXiv preprint, arXiv: 1801.06146, 2018

[12] Radford A, Narasimhan K, Salimans T, et al. Improving language understanding by generative pre-training [OL]. [2020-07-09]. https://cdn. openai. com/research-covers/language-unsupervised/language_understanding_paper.pdf

[13] Devlin J, Chang Mingwei, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding [J]. arXiv preprint, arXiv: 1810.04805, 2019

[14] Zhang Zhengyan, Han Xu, Liu Zhiyuan, et al. ERNIE: Enhanced language representation with informative entities [J]. arXiv preprint, arXiv: 1905.07129, 2019

[15] Wang Yu, SunYining, Ma Zuchang, et al. An ERNIE-based joint model for Chinese named entity recognition [J]. Applied Sciences, 2020, 10(16): No.5711

[16] Xiao Dongling, Zhang Han, Li Yukun, et al. ERNIE-GEN: An enhanced multi-flow pre-training and fine-tuning framework for natural language generation [J]. arXiv preprint, arXiv: 2001.11314, 2020

[17] Kim Y. Convolutional neural networks for sentence classification [J]. arXiv preprint, arXiv: 1408.5882, 2014

[18] Anderson M, Anderson S L, Armen C. Towards machine ethics [C] //Proc of AAAI 2004 Workshop on Agent Organizations: Theory and Practice. Palo Alto, CA: AAAI, 2004: 1-7

[19] Anderson M, Anderson S L, Armen C. MedEthEx: A prototype medical ethics advisor [C] //Proc of the National Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2006, 21(2): 1759

[20] Arkin R C. Governing lethal behavior: Embedding ethics in a hybrid deliberative/reactive robot architecture part I: Motivation and philosophy [C] //Proc of the 3rd ACM/IEEE Int Conf on Human Robot Interaction. New York: ACM, 2008: 121-128

[21] Ashley K D, McLaren B M. Reasoning with reasons in case-based comparisons [C] //Proc of the Int Conf on Case-based Reasoning. Berlin: Springer, 1995: 133-144

[22] McLaren B M, Ashley K D. Context sensitive case comparisons in practical ethics: reasoning about reasons [C] //Proc of the 5th Int Conf on Artificial Intelligence and Law. New York: ACM, 1995: 316-325

[23] Dehghani M, Tomai E, Forbus K D, et al. An integrated reasoning approach to moral decision-making [C] //Proc of the 23rd National Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2008: 1280-1286

[24] Blass J A, Forbus K D. Moral decision-making by analogy: generalizations versus exemplars [C] //Proc of the 29th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2015: 501-507

[25] Abel D, Macglashan J, Littman M L, et al. Reinforcement learning as a framework for ethical decision making [C] //Proc of AAAI Workshop: AI, Ethics, and Society. Palo Alto, CA: AAAI, 2016: 54-61

[26] Wu Y H, Lin S D. A low-cost ethics shaping approach for designing reinforcement learning agents [J]. arXiv preprint, arXiv: 1712.04172, 2017

[27] Riedl M O, Harrison B. Using stories to teach human values to artificial agents [C] //Proc of AAAI Workshop: AI, Ethics, and Society. Palo Alto, CA: AAAI, 2016: 1-8

[28] Jentzsch S, Schramowski P, Rothkopf C, et al. Semantics derived automatically from language corpora contain human-like moral choices [C] //Proc of the 2019 AAAI/ACM Conf on AI, Ethics, and Society. New York: ACM, 2019: 37-44

[29] Schramowski P, Turan C, Jentzsch S, et al. BERT has a moral compass: Improvements of ethical and moral values of machines [J]. arXiv preprint, arXiv: 1912.05238, 2019

[30] Ziegler D M, Stiennon N, Wu J, et al. Fine-Tuning language models from human preferences [J]. arXiv preprint, arXiv: 1909.08593, 2019

[31] Frazier S, Nahian M S A, Riedl M, et al. Learning norms from stories: A prior for value aligned agents [J]. arXiv preprint, arXiv:1912.03553, 2019

[32] Gips J. Toward the ethical robot [M]. Cambridge, MA: MIT Press, 1995: 243-252

[33] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [C] //Advances in Neural Information Processing Systems. New York: NIPS, 2017: 5998-6008

[34] Liu Gang, Guo Jiabao. Bidirectional LSTM with attention mechanism and convolutional layer for text classification [J]. Neurocomputing, 2019, 337(14): 325-338

[35] Lai Siwei, Xu Liheng, Liu Kang, et al. Recurrent convolutional neural networks for text classification [C] // Proc of the 29th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2015: 2267-2273

[36] Zhou Peng, Shi Wei, Tian Jun, et al. Attention-based bidirectional long short-term memory networks for relation classification [C] //Proc of the 54th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2016: 207-212

[37] Johnson R, Zhang Tong. Deep pyramid convolutional neural networks for text categorization [C] //Proc of the 55th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2017: 562-570

[38] Safaya A, Abdullatif M, Yuret D. Kuisail at semeval-2020 task 12: BERT-CNN for offensive speech identification in social media [J]. arXiv preprint, arXiv: 2007.13184, 2020

Gu Tianlong, born in 1964. PhD, professor, senior member of CCF. His main research interests include formal methods, trustworthy artificial intelligence, ethically aligned machine design, and artificial intelligence ethics.

古天龙, 1964 年生.博士,教授,CCF 高级会员.主要研究方向为形式化方法、可信人工智能、伦理智能体设计、人工智能伦理.

Feng Xuan, born in 1996. Master candidate, student member of CCF. His main research interests include deep learning, natural language processing, and artificial intelligence ethics.

冯　旋, 1996 年生.硕士研究生,CCF 学生会员.主要研究方向为深度学习、自然语言处理和人工智能伦理.

Li Long, born in 1989. PhD, lecturer, member of CCF. His main research interests include artificial intelligence security and logic programming.

李　龙, 1989 年生.博士,讲师,CCF 会员.主要研究方向为人工智能安全和逻辑程序设计.

Bao Xuguang, born in 1985. PhD, lecturer, member of CCF. His main research interests include spatial data mining, knowledge engineering, machine learning, and human-computer interaction.

包旭光,1985 年生.博士,讲师,CCF 会员.主要研究方向为空间数据挖掘、知识工程、机器学习和人机交互.

Li Yunhui, born in 1979. PhD candidate, student member of CCF. Her main research interests include data crowdsourcing and machine learning.

李云辉,1979 年生.博士研究生,CCF 学生会员.主要研究方向为数据众包和机器学习.