网络欺凌检测综述

宋宇琦[1,2],高 旻[1,2],李骏东[3],荣文戈[4],熊庆宇[1,2]
(1. 信息物理社会可信服务计算教育部重点实验室,重庆 400044; 2. 重庆大学大数据与软件学院,重庆 400044;
3. 弗吉尼亚大学电子与计算机工程学院,夏洛茨维尔 22904; 4. 北京航空航天大学计算机学院,北京 100191)

摘 要: 网络欺凌在社交媒体平台的日益泛滥引起了研究者的广泛关注,社会科学和计算机科学研究者从不同的角度对该问题进行了研究与探讨. 为梳理这些研究,本论文对社会科学领域和计算机领域在网络欺凌方面的研究进行了调查分析. 首先概述了网络欺凌的基本研究内容和网络欺凌特征,重点讨论了各种用于网络欺凌检测的机器学习方法,包括基于监督学习、基于弱监督学习、基于预设规则和深度学习算法,随后总结了 12 个现有的网络欺凌检测数据集和常用的检测性能评价指标,最后对基于异构信息网络、融合多种辅助信息和结合心理学特征的欺凌检测方法等进行了展望.

关键词: 网络欺凌; 欺凌检测; 机器学习; 社交网络; 欺凌特征; 网络欺凌数据集
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112 (2020)06-1220-10
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2020.06.025

A Survey of Cyberbullying Detection

SONG Yu-qi[1,2], GAO Min[1,2], LI Jun-dong[3], RONG Wen-ge[4], XIONG Qing-yu[1,2]
(1. *Key Laboratory of Dependable Service Computing in Cyber Physical Society, Ministry of Education, Chongqing* 400044, *China*;
2. *School of Big Data & Software Engineering, Chongqing University, Chongqing* 400044, *China*;
3. *Department of Electrical and Computer Engineering, University of Virginia, VA* 22904, *US*;
4. *School of Computer Science and Engineering, Beihang University, Beijing* 100191, *China*)

Abstract: Cyberbullying has attracted the increasing attention among researchers. Social and computer science researchers have explored cyberbullying from various perspectives. This paper surveys the existing work on cyberbullying detection in social and computer science domains. It first introduces the basic research problems and characteristics of cyberbullying; second, it discusses a variety of machine learning algorithms for cyberbullying detection, including supervised learning, weakly supervised learning, rule-based and deep learning algorithms; and third, it summarizes 12 existing datasets used in cyberbullying detection and the popular metrics for detection performance. Finally, the paper analyzes the potential research from several aspects, such as cyberbullying detection approaches based on heterogeneous information network, auxiliary information fusion, and psychological characteristics.

Key words: cyberbullying; bullying detection; machine learning; social network; bullying feature; cyberbullying dataset

1 引言

目前,网络欺凌在社交媒体中的日益泛滥以及对青少年的严重影响使其受到研究者的广泛关注[1,2]. 网络欺凌会造成受害者精神错乱,甚至严重影响身心健康,遭受过网络欺凌的青少年比没有经历过网络欺凌的青少年的自杀意念高 3.12 倍[3],网络欺凌的治理已迫在眉睫.

传统欺凌是指一个或一群人故意且重复地对其他人实施负面行为,伤害或使其不适[4]. 网络欺凌是指任何个体或群体通过电子设备或数字媒体进行实施,重复地传达出敌对或侵略的信息,旨在对他人造成伤害或不适的行为[5].

网络欺凌和传统欺凌在动机上有很大相似性,即

收稿日期:2019-01-02;修回日期:2019-11-21;责任编辑:覃怀银
基金项目: 国家自然科学基金 (No. 71102065);中央高校基金 (No. 2019CDXYRJ0011);国家重点研发计划 (No. 2018YFF0214706);广西科技重大专项(No. GKAA17129002)

用重复行为对其目标造成伤害,但网络欺凌利用互联网技术能在任何时间任何地点对目标造成伤害,因此其将受害范围扩大到校园之外. 网络欺凌还具有匿名性和非及时反馈性,可以降低承担风险的后果,减轻欺凌者的心理负担,导致网络欺凌的发生更为频繁.

从相关研究来看,网络欺凌的相关研究已深入到社会学、计算机、法律、医学等多个学科. 根据 Springer 数据库中传统欺凌与网络欺凌相关论文的学科分布来看,心理学所占比例最大,计算机科学在传统欺凌领域的研究仅占 3%,在网络欺凌领域所占比例达 12.6%.

心理学和教育学等对网络欺凌的研究集中在通过调查欺凌者的个性特质,分析人格与欺凌行为之间的关系. 通过人工检测欺凌耗时费力,且难以进行即时检测,因此针对网络欺凌的自动检测方法受到广泛关注,本文将对此方面研究进行综述.

2　网络欺凌研究内容及特征

本节将分别介绍心理学领域和计算机自动检测领域对网络欺凌的研究内容和探究的特征.

2.1　心理学家对网络欺凌的研究内容

网络欺凌虽然发生在网络上,但本质上是现实人际关系的异化,是一个复杂的社会生态系统的相互作用产生的结果[6],很多心理学的研究者对网络欺凌成因及对策有深入的研究.

目前在心理学领域主要的研究工作包括:研究多种测量网络欺凌发生频率的调查问卷,并通过抽样调查方法[7],配合访谈法来获取数据信息,以探究网络欺凌的现状和原因,分析网络欺凌的表现形式和基本特征;利用回归分析等方法探讨人口统计学信息、人格特征、行为特征等与网络欺凌之间的相关性以及其对网络欺凌被害的影响程度,揭示网络欺凌问题严峻的现状,结合实际情况,提出一系列教育等方面的科学预防及干预措施.

2.2　心理学领域研究的特征

在心理学领域对网络欺凌者的研究可以分为对人口统计学特征、人格特征、行为特征和外部特征等四个大类的调查分析.

2.2.1　人口统计学特征

人口统计学特征主要包括年龄和性别. 多项研究发现网络欺凌者实施网络欺凌的频率首先随年龄增长而增大,之后再逐渐降低[5],欺凌者的年龄主要集中在 10 ~ 15 岁.

网络欺凌的性别差异是复杂的[6],研究者们的研究结果不完全一致. 有研究认为男性用户比女性用户实施更多的网络欺凌行为,而女性用户越来越多的成为被欺凌的对象[8]. 另有研究者认为网络欺凌是一种间接欺凌,女性比男性的参与度高[9],还有研究表示网络欺凌不存在显著性别差异[10].

2.2.2　人格特征

人格特征也是影响网络欺凌的重要因素,如同情心、自尊、大五人格、黑暗三人格、自我控制、道德认同和道德推脱等特征.

富有同情心的人倾向于理解他人,研究发现缺乏同情心的人做出网络欺凌行为的概率高于富有同情心的人[11].

自尊心低的人对他人的言论更敏感,防御性强烈,容易表现出高攻击性来实施欺凌行为;且欺凌与受欺凌都与自尊呈负相关,经常遭受欺凌会严重削弱自尊心[12].

大五人格[13]包括外倾性、宜人性、尽责性、神经质和开放性,每个维度都是两极的,例如宜人性的两极是有同情心且亲切和残酷并冷淡. 研究表明宜人性、尽责性得分高的人更富有同情心和责任感,不容易与他人起冲突,做出网络欺凌行为的可能性较低,而具有低尽责性和低宜人性人格的用户实施网络欺凌的概率较高.

黑暗三人格主要研究人格中的黑暗面,包括马基雅维利主义、精神病态和自恋,研究发现这三个要素都和网络欺凌呈现显著的正相关性[14].

自我控制需要人们有意识地控制冲动,并且具有抵制利益诱惑的能力. 有研究表明人们的自我控制能力越低,造成欺凌的可能性越高[15].

道德推脱是指个体通过一定方式重新解释自身行为使其合理化,从而避免自我制裁的一种认知倾向,这种认知倾向能够减少自己在行为后果中的责任[15]. 一种观点认为随着道德标准降低,个体的道德机制容易失去自我调节作用,产生道德推脱;也有观点认为道德推脱已经没有发挥作用的必要,个体不需要道德推脱也会产生欺凌行为.

道德认同是个体围绕一套道德特质而组织起来的自我认识. 其在维持个体的道德形象上发挥着自我调节作用[16],与网络欺凌呈负相关关系.

2.2.3　行为特征

行为特征是指人类对刺激所做出的可观察和测量到的反应特点,研究者根据用户在网络环境中的行为特点分析其与网络欺凌的关系.

文献[17]以我国 1438 名高中生为研究样本进行调查,证明了行为特征会造成实施或受到网络欺凌的概率差异. 例如,上网时间越多,产生网络冲突的概率越高,网络欺凌行为发生的频率越高. 拥有的网络通讯平台数量越多,上网的机会也就越大,参与网络欺凌的概率越大. 学生成绩的好坏也与网络欺凌发生的概率相关:学习成绩较差的学生容易沉迷于网络,更容易成为网络欺凌的欺凌者.

2.2.4 外部特征

外部特征主要指家庭环境. 有研究表示,当父母长辈等对青少年上网活动干预越多、管理越严格,青少年对网络的接触时间则较少,其实施网络欺凌行为的概率就越小. 而当家庭不和睦、亲子关系疏远时,孩子受到家长的管教约束及正确引导较少,参与网络欺凌的可能性则相对较高[18].

2.3 计算机自动检测的研究内容

计算机自动检测主要判断一条评论是否属于网络欺凌文本,也有借助图片等信息检测网络欺凌行为[19],但主要集中在文本内容的检测.

在社交网络平台的海量信息中识别欺凌文本内容是网络欺凌检测中的一项基本任务,这一研究可以视为二分类任务,通常使用文本分类、主题分类、情感分析等技术对消息、发送者和接收者的特征进行处理,从而检测网络欺凌行为[20].

2.4 计算机自动检测网络欺凌的特征

在自动检测网络欺凌研究中,主要使用的特征包括文本特征、情感特征、用户特征和网络特征.

2.4.1 文本特征

文本特征是使用最广泛的特征,例如通过自然语言处理方法提取的亵渎语、代词、关键词、N 元语言模型、词袋模型(Bags-of-Words)、词频-逆文档频率等特征[21~24].

大多数网络欺凌信息具有侮辱性词汇,这些词可以从亵渎词词典(https://www. noswearing. com/)或外部公开词典(https://www. urbandictionary. com/)获取. 但并非只通过亵渎词就能确认欺凌行为,包含人称或人称代词的亵渎短语更可能反映欺凌行为[25].

关键词常与种族、外貌、性别等主题相关[26]. 研究者创建与网络欺凌高度相关的关键词词汇表,筛选出可能是欺凌的文本,如“丑”、“肥胖”.

N 元语言模型是根据前 $N-1$ 个词预测第 N 个词出现的概率[27]. BoW 模型忽略文本中单词的顺序及语法,只将其视为单词的汇合,且每个单词独立出现[28]. 词频-逆文档频率用来评估字/词对于语料库中一份文本的重要程度[29,30]. 字/词的重要性与在文本中出现的次数成正比,与在语料库中出现的频率成反比. 以上三种模型通常与其他特征结合使用以提高检测网络欺凌的准确率.

此外,还有文档长度、特殊字符、词性标注、计算字符串相似度[31]等其他特征.

2.4.2 情感特征

情感特征是指用户言语中所表达出的情感倾向. 研究者们利用情感关键词、短语和表情符号对语料库进行情感分析,抽取特征并判断情感倾向,从而判断是否属于网络欺凌行为.

由于欺凌文本中包含上述情绪的文本只占 6%,而某些情绪又带有玩笑性质,仅凭短文本难以区分欺凌和玩笑,因此情感特征常需要与其他方法结合进行检测. 例如刘欢[3]等使用 TF-IDF 模型提取内容特征、使用 K 近邻图挖掘情感相似性、融入用户和推文关系,三者联合学习进行检测.

2.4.3 用户特征

用户特征是指利用用户个人的信息进行判定,包括年龄、性别、性取向、种族和行为等.

研究发现,纳入性别特征后显著改善了检测算法的准确性[32]. 但是,由于系统中用户年龄和性别等信息缺乏真实性,上述方法需要额外模块验证信息的准确性以及从更多途径捕获合法的用户特征.

2.4.4 网络特征

网络特征是指社交网络中用户的信息,诸如发表推文数、推文包含链接数、话题数#、推文发表频率等. 有研究通过网络特征计算用户在社交网络中的活跃程度来检测 Twitter 中的欺凌行为[33].

目前网络特征主要包含数据特征,例如粉丝数量、发文频率、视频上传数量等,没有充分探索和利用用户的关联关系.

当前多数研究都侧重于文本特征而在情感特征、用户特征和网络特征方面的研究相对较少,同时研究四大类特征的更是少数.

2.5 心理学研究的特征与自动检测特征的对比

以上两节分别介绍了在心理学领域对网络欺凌行为研究分析的特征以及计算机自动检测网络欺凌使用的特征(汇总在图 1),心理学对网络欺凌的研究大多集中在对人格特征的探索,计算机自动检测工作主要集中在文本和网络特征的研究,共同的特征目前仅有年龄、性别和上网时长.

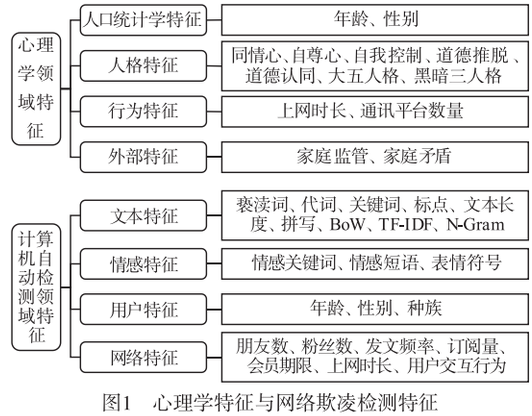


图1 心理学特征与网络欺凌检测特征

3 网络欺凌检测方法

网络欺凌检测算法流程如图 2 所示. 流程第一阶

段:特征工程——从数据集中抽取有代表性和有区分度的特征. 第二阶段:拟合分布——学习数据集中通过第一阶段提取特征的分布规律形成分类器. 第三阶段:预测标签——根据第二阶段训练的分类器判别待检测记录,确定其是否属于网络欺凌行为.

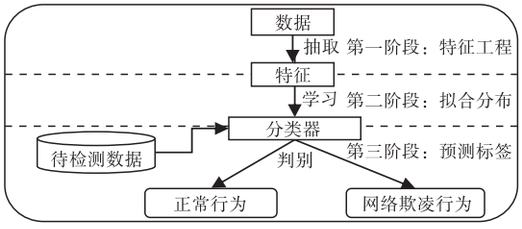


图2　网络欺凌检测算法基本流程

3.1　根据学习方法的分类

根据学习方法,网络欺凌检测方法分为基于监督学习的检测方法和基于弱监督学习的检测方法.

3.1.1　基于监督学习的检测方法

基于监督学习的方法是现阶段研究最多的方法. 早在 2009 年,研究者就从社交网络数据集分析发现,欺凌文本和其相邻的正常文本有明显的差异[29],提出通过评论相似度的检测方法.

文献[21]将评论按照网络欺凌主题进行聚类,在每个主题内学习特征,再结合文本中的亵渎词和消极情绪进行检测.

文献[30]提出了两个新特征提取方法:使用人称代词计数和利用 Skip-gram 构建特征向量,并将长距离单词作为特征;检测负面性更强的攻击评论.

有研究者提出新的表示学习框架——增强的词袋模型(EBoW)[34](如图 3),首先扩展预定义的亵渎词列表并分配不同的权重学习网络欺凌特征,再将词袋特征、潜在语义特征和欺凌特征联合进行表示学习,之后将特征向量输入分类器进行训练.

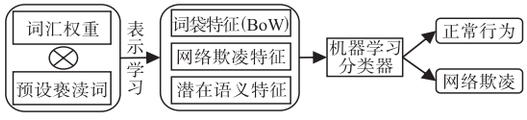


图3　EBoW表示学习框架图

另有研究者提出利用网络拓扑结构特征区分正常用户、网络欺凌者和受害者的检测方法[35]. 通过分析网络中节点度、平均邻居度、嵌入性、公共邻居数目、节点偏好连接等特征,将特征分组成对,不符合正常行为模式的为网络欺凌者和受害者.

大多数监督学习方法需要提取特征,但不同数据的特征是不同的,提取特征需要消耗大量人力.

3.1.2　基于弱监督学习的检测方法

为解决数据集中的标签信息严重不均衡的问题,有研究者探究了基于弱监督学习的检测方法. 例如,利用专家提供的标示欺凌行为的种子词汇表,在大规模无标签社交媒体交互语料库中提取指示网络欺凌的其他词汇[36].

另有研究通过最大化-参与者-词汇一致性这一指标,评估网络欺凌行为,同时考虑社会结构信息,推断容易欺负他人和成为受害者的用户[37].

3.2　根据预设规则划分的方法

预设规则是指遵循一系列预先设定的词典或规则,判别检测对象是否为欺凌行为.

3.2.1　基于词典的检测方法

基于词典的检测方法,首先建立亵渎词词典,再根据文本包含亵渎词的多少判断是否存在欺凌行为. 例如,通过 noswearing 网站编制亵渎词词典,从 formspring.me 问答网站搜集包含词典中任意亵渎词的文本,每条文本根据包含亵渎词数量的多少打分,再结合有监督的学习方法检测欺凌消息[23].

也有研究将文本通过词法分析器进行处理,并与内容分析模块中建立的欺凌攻击模型进行比较[22],并使用颜色表示欺凌不同等级. 消息发送者对应的欺凌等级根据每条消息更新. 系统根据用户的颜色状态对其做出干预或限制.

3.2.2　基于规则的检测方法

基于规则的检测需要预先设置规则进行判别. 例如,研究[38]对数据按照年龄、在线时间、在线活动是否存在风险进行分类,将不同类别之间的关系与受网络欺凌风险设置相关规则,根据输入的数据匹配对应的规则进行分类,以检测处于不同网络欺凌风险中的青少年. 研究框架如图 4,其中规则 A2&B1&C1 = V 是指 13 ~ 16 岁的青少年每周在线时间超过 8 小时并经常参与社交网络活动,其遭受网络欺凌的风险非常高.

还有研究者结合历史对话和写作风格等特征设计基于词法和句法的框架检测欺凌文本和欺凌用户[24],使用人工设计的规则计算 YouTube 评论中句子的欺凌分数;还有研究者通过规则识别亵渎词和人物指示之间关系的模式,通过人称代词、姓名以及“@”行为来挖掘用户的指示关系. 如果一条推文匹配到至少一条规则,就将它归为欺凌.

3.2.3　基于混合检测的方法

混合检测方法指在检测的过程中用人类掌握的知识进行推理. 有研究提出如图 5 所示的两种混合方法[25]. 方法(1)邀请专家为用户基础特征提供权重,再将权重提供给系统加权计算每个用户的欺凌分数,结合原始特征和专家系统评判的特征进行训练. 方法(2)则先训练分类器,再将其与专家系统结合成一个混合

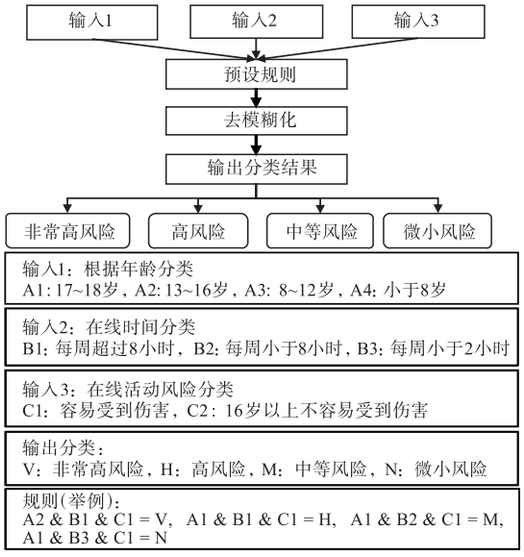


图4 基于规则的检测方法框架示例图

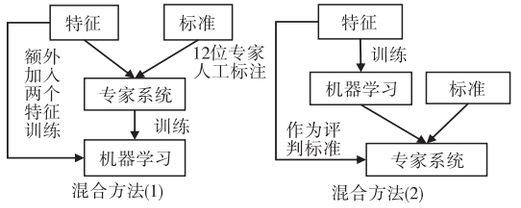


图5 基于专家系统和机器学习的混合检测方法

系统.实验结果显示通过混合方法进行检测的效果优于两个子方法分别检测的结果,同时混合方法(2)的性能优于混合方法(1).

3.3 根据分类器划分的方法

传统的机器学习检测算法具有坚实的理论与研究基础,随着深度学习的不断发展,已有研究将深度学习方法应用于检测网络欺凌.

3.3.1 传统机器学习检测方法

使用传统机器学习进行检测的有基于朴素贝叶斯[21]、决策树[25]、支持向量机[29,32]、逻辑回归[30]和随机森林[33]等检测方法.

3.3.2 深度学习检测方法

大多数检测依赖特征选择,同一特征在不同数据集的检测性能不够稳定,不合适的特征会降低检测精度.近几年,有研究将深度学习运用到网络欺凌检测中,减少对显示特征的依赖,提高检测精度.

例如有研究者提出语义增强的边缘化降噪自动编码器(smSDA)[28],从正常词汇中学习欺凌特征的潜在结构,它不需要根据欺凌词汇来检测欺凌消息,这将解决欺凌文本中不包含欺凌词汇的问题.

在文献[39]中,使用递归神经网络对文本特征和社交网络元数据特征建模学习,用连接层将其合并,以进行分类检测,算法框架如图6所示.该方法首次在网络欺凌检测领域提出利用元数据作为辅助信息来补全文本特征,展示了多分类器的结合和元信息的融合在检测方面的优势.

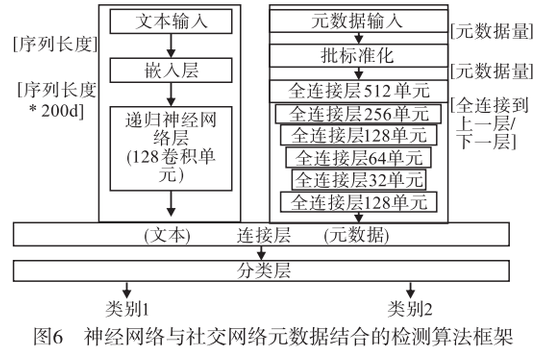


图6 神经网络与社交网络元数据结合的检测算法框架

另有研究者提出了一种基于长短期记忆网络(LSTM)的网络欺凌检测方法[40],这种方法在历史信息中对每个用户进行情感倾向分析,同时基于单词出现的频率将输入的单词数据矢量化,然后利用基于LSTM的分类器对这些信息进行处理,输出两类结果:投票机制决策和信心机制决策的权重,最后加权这两种决策结果以多数表决的方式作为最终分类结果的输出.

近两年,还有研究利用Instagram社交网络中的图像和图像标题训练卷积神经网络,以协助对网络欺凌评论的检测[19];有研究采用卷积神经网络和增加的特征密度检测网络欺凌[41];还有工作使用卷积神经网络、双向长短期记忆网络和有关联机制的双向长短期记忆网络模型并结合迁移学习实现跨社交平台的网络欺凌检测[42].

3.3.3 其他方法

除上述检测方法外,还有一些无法根据上述规则进行分类的方法.例如,有研究向虚拟的环境中添加规范的代理监控用户在虚拟世界的行为,将检测问题描述为一种违反规范的问题[43].

本文对具有代表性的网络欺凌检测方法分类如表1所示,目前主流的研究是使用监督学习和机器学习的方法,弱监督学习可以解决部分对数据标签依赖性的问题,基于预设规则的研究逐渐减少,结合深度学习进行欺凌检测的方法在逐渐增多.

为了展现网络欺凌检测算法的检测性能,本文选取了部分典型研究及其使用的数据集、评价指标和实验数值结果汇总在表2.同时也在该表中展示了相关研究者在其论文中所属的研究机构,有利于感兴趣的读者追踪其后续发展.

表 1　部分网络欺凌检测方法所属类别

研究	监督学习	弱监督学习	基于词典	基于规则	混合方法	机器学习	深度学习	其他方法
Yin,et al. ,2009[29]	✓					✓		
Dinakar,et al. ,2011[21]	✓					✓		
Serra and Venter,2011[38]				✓				
Bosse and Stam,2011[43]								✓
Pérez,et al. ,2012[22]			✓					
Kontostathis,et al. ,2013[23]	✓		✓			✓		
Bretschneider,et al. ,2014[24]				✓				
Dadvar,et al. ,2014[25]					✓	✓		
Chavan and Shylaja,2015[30]	✓					✓		
Al-garadi,et al. ,2016[33]	✓					✓		
Zhao,et al. ,2016[34]	✓						✓	
Zhong,et al. ,2016[19]							✓	
Ptaszynski,et al. ,2017[41]							✓	
Zhao and Mao,2017[28]	✓						✓	
Ptaszynski,et al. ,2017[41]							✓	
Agrawal and Awekar,2018[42]							✓	
Pitsilis,et al. ,2018[40]							✓	

表 2　部分网络欺凌检测方法总结

研究	研究机构	数据集	评价指标	实验结果
Yin,et al. 2009[29]	Lehigh University	MySpace	F1	0.313
Dinakar,et al. ,2011[21]	MIT	YouTube	Accuracy	0.667
Bretschneider,et al. ,2014[24]	Martin-Luther-University	Twitter	F1	0.719
Chavan,et al. ,2015[30]	Visvesvaraya Technological University	Kaggle	Precision	0.769
Al-garadi,et al. ,2016[33]	University of Malaya	Twitter	F1 值	0.941
Zhao,et al. ,2016[34]	Nanyang Technological University	Twitter	F1 值	0.78
Zhong,et al. ,2016[19]	Pennsylvania State University	Instagram	Accuracy	0.95
Ptaszynski,et al. ,2017[41]	Kitami Institute of Technology	MeCab	Precision	0.936
Zhao,et al. ,2017[28]	Nanyang Technological University	MySpace	F1 值	0.776
Agrawal,et al. ,2018[42]	Indian Institute of Technology	Wikipedia	F1 值	0.94

4　网络欺凌检测数据集与检测指标

数据集是网络欺凌检测效果的验证基础,检测指标是评估算法性能的度量标准,目前广泛使用的网络欺凌数据集是从社交网站爬取的评论、问答、发帖回复、图片和视频等信息,公开数据集的下载链接已汇总在 GitHub 平台(https://github.com/ CQU-CSE/Dataset-Collection).

4.1　数据集

文献[44]从 Formspring 网站爬取 18554 位用户的问答数据,随机抽取 3915 条消息通过亚马逊外包服务进行人工标记,其中 369 条被标记为欺凌.

文献[45]对 MySpace 网站数据进行收集,包含用户信息(性别、年龄、城市、省、市)和对话文本信息. 其中小部分数据由研究助手投票标记类别,在 2088 条数据中标记为欺凌行为的有 434 条.

文献[46]爬取了 Ask.fm 网站中的部分用户信息和问答内容,目前该数据集没有公开的标注信息.

有研究者收集 Instagram 网站中用户信息、图片信息及相关评论,并通过众包平台进行标注,1954 条数据中有 567 条是网络欺凌行为[47].

还有研究人员通过观看 Vine 网站 6 秒的视频及其

评论,对相关行为进行标注[48],在 971 条数据中有 304 条为网络欺凌.

BullyingV3.0[49]是从 Twitter 网站爬取的包含 7321 条推文的数据集,研究人员对文本分类且标记了用户在网络欺凌中的角色和情感标签,其中 2101 条为网络欺凌.

WOW 和 LOL[50]数据集分别来自魔兽论坛和英雄联盟论坛,标签由三位专家人工标记. WOW 有 16975 条信息,137 条是欺凌信息,LOL 包含 17354 条信息,207 条是欺凌信息.

文献[51]从 Twitter 中爬取 1303 位用户和 9484 条评论信息并通过众包平台对用户进行标注,有 58 名用户被标记为网络欺凌者.

文献[27]根据 Wikipedia 讨论区的评论生成 6 千万条评论的语料库,从语料库中随机抽样 37611 条记录,根据多数投票标记,其中欺凌行为达到 0.9%.

Harassment-Corpus[52]包括一个骚扰词语料库和从 Twitter 社交平台爬取的评论数据集. 语料库共 725 个单词并根据性别、种族、外貌和政治等特征分为 6 类;数据集通过人工对 24189 条记录进行标记,有 3119 条标记为网络欺凌行为.

Hate and Abusive Speech[53]数据集也是从 Twitter 中爬取的用户及评论数据,通过众包平台对 99799 条数据标注,其中 46009 条为欺凌行为.

4.2 网络欺凌检测指标

网络欺凌检测被视为二分类任务,即将行为划分为正常行为和网络欺凌行为两类,因此检测指标使用分类常用的正确率、准确率、召回率和 F1 值.

正确率指被正确预测类别的行为数量占所有待预测行为数量的比例. 准确率指正确地被预测出的网络欺凌行为数量占所有预测为网络欺凌数量的比例. 召回率指预测出的网络欺凌行为占检测样本集中所有网络欺凌行为的比例. F1 值综合准确率和召回率两个指标体现算法的整体性能.

5 网络欺凌检测研究展望

现有的研究工作为网络欺凌检测奠定了坚实的基础,但仍存在一些亟需解决的问题,本节将对网络欺凌检测的研究方向做以展望.

5.1 欺凌检测数据扩展

当前欺凌检测数据严重依赖众包平台或者专家人工标注,阻碍了有监督检测方法的发展. 针对这一问题,可以探究将弱监督学习与深度学习结合的方法. 例如通过生成式对抗网络学习已有的标签数据,模拟生成符合网络欺凌特征的数据,以增加欺凌样本的数据量,提高检测性能.

5.2 基于异构信息网络的检测

目前大多数检测方法依赖文本信息,但文本呈现出越来越明显的长度短、噪声多、无结构和故意混淆等特点;并且同时存在多种语言也给检测网络欺凌带来了新的挑战[1]. 针对这一问题,可以利用异构信息网络与词嵌入结合的研究方法,通过用户、博文、评论及用户行为构建异构信息网络学习节点嵌入信息后再结合分类器进行检测.

5.3 结合用户交互信息的检测

现阶段关于网络欺凌的检测研究主要集中在判别每条评论是否属于欺凌行为. 由于网络欺凌具有重复性,仅凭单条文本信息来判定其是否属于欺凌有失严谨. 因此如何收集用户互动记录的数据集、如何充分利用信息的上下文环境检测出具有重复性的网络欺凌行为是现阶段面临的一大挑战.

5.4 融合多种辅助信息的检测

随着网络欺凌形式的多样化,已有研究开始结合图片与文本信息以提高网络欺凌检测准确性[54]. 未来的研究可以结合更多辅助信息,如图片、语音、表情包、视频等内容提取多个方面的特征进行分类.

5.5 结合心理学特征的检测

心理学研究表明,人格特征是影响网络欺凌的重要因素,而网络欺凌自动检测使用的特征大多数是由人工设计的、主要关注于文本信息及其表达出的情感特征,缺乏对用户心理、人格特征的判定. 如何从用户的行为中分析其人格特征,融入人格特征进行网络欺凌检测极具现实意义.

6 总结

本文介绍了网络欺凌的概念与特征,阐明网络欺凌的危害性和监控难度,总结心理学和计算机自动检测使用的特征,分类介绍网络欺凌检测方法,汇总了网络欺凌检测数据集和常用检测指标. 最后对网络欺凌检测未来的研究方向进行了展望.

参考文献

[1] Semiu Salawu, Yulan He, Joanna Lumsden. Approaches to automated detection of cyberbullying: A survey[J]. IEEE Transactions on Affective Computing, 2017, 99(10): 1-10.

[2] Cheng L, Li J, Silva Y, et al. PI-bully: personalized cyberbullying detection with peer influence[A]. The International Joint Conference on Artificial Intelligence[C]. Palo Alto: AAAI, 2019. 5829-5835.

[3] Dani H, Li J, Liu H. Sentiment informed cyberbullying detection in social media[A]. The European Conference on

Machine Learning and Principles and Practice of Knowledge Discovery in Databases[C]. Cham:Springer,2017. 52 –67.

[4] Olweus D. Aggressive Behavior ‖ Bullying at School[M]. Boston,MA:Springer,1994. 97 –130.

[5] Tokunaga R S. Following you home from school:a critical review and synthesis of research on cyberbullying victimization[J]. Computersin Human Behavior,2010,26(3):277 –287.

[6] 杨坤玉. 青少年网络欺凌行为介入模式的 GAM 探索[D]. 兰州:兰州大学哲学,2017.

[7] Wong R Y M,Cheung C M K,Xiao B. Does gender matter in cyberbullying perpetration? An empirical investigation[J]. Computers in Human Behavior,2018,79(2):247 –257.

[8] Schneider S K,O'donnell L,Stueve A,et al. Cyberbullying,school bullying,and psychological distress:a regional census of high school students[J]. American Journal of Public Health,2012,102(1):171 –177.

[9] Kowalski R M,Limber S P. Electronic bullying among middle school students[J]. Journal of Adolescent Health,2007,41(6):S22 –S30.

[10] Slonje R,Smith P K. Cyberbullying:another main type of bullying? [J]. Scandinavian Journal of Psychology,2008,49(2):147 –154.

[11] Brewer G,Kerslake J. Cyberbullying,self-esteem,empathy and loneliness[J]. Computers in Human Behavior,2015,48(1):255 –260.

[12] 刘琳. 中学生传统欺凌、网络欺凌及其与自尊的关系[D]. 沈阳:沈阳师范大学,2014.

[13] 苑广哲. 大学生大五人格和网络欺凌行为的关系:自我控制的中介作用[D]. 南京:南京师范大学,2017.

[14] Goodboy A K,Martin M M. The personality profile of a cyberbully:examining the dark triad[J]. Computers in Human Behavior,2015,49(4):1 –4.

[15] 杨继平,王兴超,高玲. 道德推脱对大学生网络偏差行为的影响:道德认同的调节作用[J]. 心理发展与教育,2015,31(3):311 –318.

[16] Winterich K P,Aquino K,Mittal V,et al. When moral identity symbolization motivates prosocial behavior:the role of recognition and moral identity internalization[J]. Journal of Applied Psychology,2013,98(5):759.

[17] Zhou Z,Tang H,Tian Y,et al. Cyberbullying and its risk factors among Chinese high school students[J]. School Psychology International,2013,34(6):630 –647.

[18] Yang X,Wang Z,Chen H,et al. Cyberbullying perpetration among Chinese adolescents:the role of interparental conflict,moral disengagement,and moral identity[J]. Children and Youth Services Review,2018,86:256 –263.

[19] Zhong H,Li H,Squicciarini A C,et al. Content-driven detection of cyberbullying on theInstagram social network[A]. The International Joint Conference on Artificial Intelligence[C]. Palo Alto:AAAI Press,2016. 3952 –3958.

[20] Haidar B,Chamoun M,Yamout F. Cyberbullying detection:a survey on multilingual techniques[A]. The European Modelling Symposium on Computer Modelling and Simulation[C]. New York:IEEE,2016. 165 –171.

[21] Dinakar K,Reichart R,Lieberman H. Modeling the detection of textual cyberbullying[A]. The International Conference on Weblogs and Social Media[C]. CA:AI Access Foundation,2011. 11 –17.

[22] Pérez P J C,Valdez C J L,Ortiz M G C,et al. MISAAC:Instant messaging tool for cyberbullying detection[A]. The International Conference on Artificial Intelligence[C]. Las Vegas:CSREA Press,2012. 1 –4.

[23] Kontostathis A,Reynolds K,Garron A,et al. Detecting cyberbullying:query terms and techniques[A]. The Annual Web Science Conference[C]. New York:ACM,2013. 195 –204.

[24] Bretschneider U,Wöhner T,Peters R. Detecting online harassment in social networks[A]. The International Conference on Information Systems[C]. Auckland:AIS,2014. 1 –14.

[25] Dadvar M,Trieschnigg D,de Jong F. Experts and machines against bullies:a hybrid approach to detect cyberbullies[A]. The Canadian Conference on Artificial Intelligence[C]. Cham:Springer,2014. 275 –281.

[26] Dadvar M,Trieschnigg R B,de Jong F M G. Expert knowledge for automatic detection of bullies in social networks[A]. The Benelux Conference on Artificial Intelligence[C]. Delft:University of Groningen Press,2013. 57 –64.

[27] Wulczyn E,Thain N,Dixon L. Ex machina:Personal attacks seen at scale[A]. The International World Wide Web Conference[C]. New York:ACM,2017. 1391 –1399.

[28] Zhao R,Mao K. Cyberbullying detection based on semantic-enhanced marginalized denoising auto-encoder[J]. IEEE Transactions on Affective Computing,2016,8(3):328 –339.

[29] Yin D,Xue Z,Hong L,et al. Detection of harassment on web 2. 0[A]. The Content Analysis in the WEB[C]. New York:ACM,2009. 1 –7.

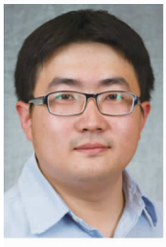
[30] Chavan V S,Shylaja S S. Machine learning approach for detection of cyber-aggressive comments by peers on social media network[A]. The International Conference on Ad-

vances in Computing, Communications and Informatics [C]. DC:IEEE,2015. 2354 – 2358.
[31] Sood S O, Antin J, Churchill E. Using crowdsourcing to improve profanity detection[A]. The AAAI Spring Symposium[C]. CA:AI Access Foundation,2012. 69 – 74.
[32] Dadvar M, Jong F M G, Ordelman R, et al. Improved cyberbullying detection using gender information[A]. The Dutch-Belgian Information Retrieval Workshop[C]. Ghent:Ghent University Press,2012. 1 – 3.
[33] Al-garadi M A, Varathan K D, Ravana S D. Cybercrime detection in online communications:the experimental case of cyberbullying detection in the twitter network[J]. Computers in Human Behavior,2016,63(C):433 – 443.
[34] Zhao R, Zhou A, Mao K. Automatic detection of cyberbullying on social networks based on bullying features[A]. The International Conference on Distributed Computing and Networking[C]. New York:ACM,2016. 43 – 48.
[35] Chelmis C, Zois D S, Yao M. Mining patterns of cyberbullying on twitter[A]. The International Conference on Data Mining Workshops[C]. New York:IEEE,2017. 126 – 133.
[36] Raisi E, Huang B. Cyberbullying detection with weakly supervised machine learning[A]. The International Conference on Advances in Social Networks Analysis and Mining[C]. New York:ACM,2017. 409 – 416.
[37] Raisi E, Huang B. Weakly supervised cyberbullying detection with participant-vocabulary consistency[J]. Social Network Analysis and Mining,2018,8(1):38 – 53.
[38] Serra S M, Venter H S. Mobile cyber-bullying:a proposal for a pre-emptive approach to risk mitigation by employing digital forensic readiness[A]. The Conference on Information Security for South Africa[C]. IEEE Computer Society,2011. 1 – 5.
[39] Founta A M, Chatzakou D, Kourtellis N, et al. A unified deep learning architecture for abuse detection[A]. The Conference on Web Science[C]. New York:ACM,2019. 105 – 114.
[40] Pitsilis G K, Ramampiaro H, Langseth H. Detecting Offensive Language in Tweets Using Deep Learning[DB/OL]. https://arxiv. org/abs/1801. 04433, 2018-08-16/2019-11-28.
[41] Ptaszynski M, Eronen J K K, Masui F. Learning deep on cyberbullying is always better than brute force[A]. The Linguistic and Cognitive Approaches to Dialog Agents Workshop[C]. Aachen:CEUR-WS,2017. 19 – 25.
[42] Agrawal S, Awekar A. Deep learning for detecting cyberbullying across multiple social media platforms[A]. The European Conference on Information Retrieval [C]. Cham:Springer,2018. 141 – 153.
[43] Bosse T, Stam S. A normative agent system to prevent cyberbullying[A]. The International Conference on Intelligent Agent Technology[C]. DC:IEEE,2011. 425 – 430.
[44] Reynolds K, Kontostathis A, Edwards L. Using machine learning to detect cyberbullying[A]. The 10th International Conference on Machine Learning and Applications [C]. New York:IEEE,2011. 241 – 244.
[45] Bayzick J, Kontostathis A, Edwards L. Detecting the presence of cyberbullying using computer software[A]. Web Science Conference[C]. New York:ACM,2011. 93 – 96.
[46] Hosseinmardi H, Ghasemianlangroodi A, Han R, et al. Towards understanding cyberbullying behavior in a semi-anonymous social network[A]. The International Conference on Advances in Social Networks Analysis and Mining[C]. DC:IEEE,2014. 244 – 252.
[47] Hosseinmardi H, Mattson S A, Rafiq R I, et al. Analyzing labeled cyberbullying incidents on the instagram social network[A]. The International Conference on Social Informatics[C]. Cham:Springer,2015. 49 – 66.
[48] Rafiq R I, Hosseinmardi H, Han R, et al. Careful what you share in six seconds:Detecting cyberbullying instances in vine[A]. The International Conference on Advances in Social Networks Analysis and Mining[C]. New York:ACM,2015. 617 – 622.
[49] Sui J. Understanding andFighting Bullying with Machine Learning[D]. USA WI:The Univ of Wisconsin-Madison,2015.
[50] Bretschneider U, Peters R. Detecting cyberbullying in online communities[A]. The European Conference on Information Systems[C]. Istanbul:AIS,2016. 61 – 74.
[51] Chatzakou D, Kourtellis N, Blackburn J, et al. Mean birds:Detecting aggression and bullying on Twitter[A]. The Web Science Conference[C]. New York:ACM,2017. 13 – 22.
[52] Rezvan M, Shekarpour S, Balasuriya L, et al. A quality type-aware annotated corpus and lexicon for harassment research[A]. The Conference on Web Science[C]. New York:ACM,2018. 33 – 36.
[53] Founta A M, Djouvas C, Chatzakou D, et al. Large scale crowdsourcing and characterization of twitter abusive behavior[A]. The AAAI Conference on Web and Social Media[C]. Palo Alto:AAAI,2018. 491 – 500.
[54] Cheng L, Li J, Silva Y N, et al. Xbully:cyberbullying detection within a multi-modal context[A]. The International Conference on Web Search and Data Mining[C]. New York:ACM,2019. 339 – 347.

作者简介

宋宇琦　女,1994 年生,2019 年 6 月毕业于重庆大学大数据与软件学院,获得硕士学位. 主要研究方向为机器学习、虚假用户检测等.
E-mail:songyq@ cqu. edu. cn

高 旻(通信作者)　女,1980 年生,工学博士,重庆大学大数据与软件学院副教授、硕士生导师. 主要研究包括推荐系统、异常检测、社会媒体挖掘.
E-mail:gaomin@ cqu. edu. cn



李骏东　男,1990 年生,工学博士,美国弗吉尼亚大学电子与计算机工程和计算机科学系助理教授. 主要研究兴趣为数据挖掘和机器学习.
E-mail:jundong@ virginia. edu

荣文戈　男,1975 年生,工学博士,北京航空航天大学计算机学院教授、博士生导师. 主要研究方向为机器学习、自然语言处理、数据挖掘和信息系统等.
E-mail:w. rong@ buaa. edu. cn

熊庆宇　男,1965 年生,工学博士,重庆大学大数据与软件学院教授、博士生导师. 2002 年 3 月毕业于日本九州大学. 主要研究方向为智能控制、传感器网络和信息系统.
E-mail:xiong03@ cqu. edu. cn