文章编号：2095-2163(2020)03-0218-05 中图分类号：TP391 文献标志码：A

面向聊天机器人的敏感内容识别研究

朱泽圻
（哈尔滨工业大学，哈尔滨 150001）

摘 要：本文提出在聊天机器人的应用背景下敏感内容的定义，统计了各种分类标准下敏感内容的分布，并在从网络爬取的问答语料中，分别采用敏感词表过滤与机器学习方法进行了数据清洗，在提出的敏感内容定义下，召回率达到 80%，合格数据留存率达到 60%。本文还利用优化后的敏感词表与启发式规则，无监督地获得敏感语料，可有效地大量获得无关键词的敏感内容，扩增数据合格率可达 80%。

关键词：聊天机器人；文本分类；敏感内容识别

Research on sensitive content recognition for chat robot

ZHU Zeqi
（Harbin Institute of Technology, Harbin 150001, China）

【Abstract】 In this paper, the definition of sensitive content as well as the statistical analysis of different kinds of sensitive content in the context of chat robot application are proposed. Sensitive vocabulary filtering and machine learning are used to clean the question-and-answer corpus crawled from the network. Under the proposed definition of sensitive content, the recall rate reaches 80%, and the qualified data retention rate reaches 60%. Unsupervised mining method with optimized sensitive vocabulary and heuristic rules are used to obtain sensitive corpus, which effectively obtains a large number of sensitive content without keywords. The eligibility rate of the expanded data can reach 80%.

【Key words】 chat robot; text classification; sensitive content recognition

0 引 言

聊天机器人是一种人机交互系统，通过自然语言模拟人类进行对话。这种系统往往运行在各种平台上，如个人电脑、社交网络软件或即时通讯工具等。聊天机器人主要有 2 种类型：封闭域聊天机器人进行的是带有目的的对话，以尽快获得必要信息、完成任务为目标；开放域聊天机器人进行的则是非任务型对话，也就是所谓的闲聊，以持续推进聊天为目标[1-2]。

时至今日，聊天机器人受到了工业界的广泛关注。聊天机器人作为人机交互问题的一种解决方案，在智能硬件等领域都陆续进入了实用，发挥了重要的作用，有着良好的商业前景。目前，各大互联网公司都先后推出了自己的聊天机器人产品，如微软的小冰、阿里的店小蜜、百度的度秘等，以聊天机器人为主打产品的创业公司也在陆续涌现。

聊天机器人产生回复的方式主要有 3 种：基于人工编写的规则，基于从问答语料库的检索和基于模型的生成。其中，检索和生成方法都对语料库有较大的需求：在检索过程中，直接在问答语料库中匹配问题，获得回答；而在生成过程中，也需要使用已有的语料库训练模型。目前的聊天机器人都非常依赖大规模语料库。

由于规模较大，聊天机器人的语料库往往是从公开网络上爬取的。然而，互联网上不仅有理性的讨论，也有不理智的辱骂、仇视与偏见。微软曾经把聊天机器人程序 Tay 上线到 twitter 上，通过与网友的互动学习对话，结果半天之内就学会了仇视人类和种族歧视的言论，引发了广泛的争论、质疑与反思[3-4]。要承担程序研发者的社会责任，就需要从语料库的构建过程开始，清洗敏感内容。

当前的敏感内容清洗手段主要目的是阻止不良信息在互联网上扩散，比较重视主题上的敏感内容[5]。但对于面向商业化应用的聊天机器人而言，除了上述明显有违国家相关法律法规的信息之外，对于可能伤害用户的内容、可能攻击其他厂商引发纠纷的内容也是不宜发表的敏感信息。此外，已有的敏感内容清洗系统往往构建静态的知识库与规则，不利于持续的扩充；然而聊天机器人系统需要持续从互联网中爬取语料，而随着时间的推移，也一定

作者简介：朱泽圻（1994-），男，硕士研究生，主要研究方向：聊天机器人的构建及其个性化。
通讯作者：朱泽圻 Email：1494011443@qq.com

收稿日期：2019-06-06

会有新的敏感内容出现，需要有扩展能力的捕捉方式。本文的目标是设计一个面向聊天机器人的敏感内容识别方案，涉及的工作包括研发一个敏感内容的清洗系统和一个敏感语料的扩增系统。

1　相关研究

1.1　敏感内容的定义

对于敏感内容的定义，以往的研究者也有多种看法。目前学界普遍认为，敏感内容分为 2 类。一类是主题上的敏感内容，另一类是态度倾向上的敏感内容[5-6]。对于主题上的敏感内容，只要识别出了主题就可以直接过滤；而对于态度倾向上的敏感内容，则需要进一步判断态度倾向。具体来说，如果一对问答提及淫秽色情的内容，那么这对问答就可以直接过滤掉；然而，如果一对问答提及的是一个犯罪事件，则需要进一步分析发言者的情感倾向、评价的对象等，最终才能决定这是不是敏感内容。然而，在聊天机器人的背景下，上述定义方式并不能完全适应需求。

迄至目前，聊天机器人的交互能力较低，表达鲜明观点的需求不高；但与此同时，一旦聊天机器人发表了不恰当的言论，除了给用户造成不适，还容易造成传播上的危机，给运营者造成不良影响。在能够通过爬虫技术得到大规模语料库的背景下，相比起查准率，更重要的是查全率。研究可知，若未能全面收录合理的对话语料不会带来太大的损失，而错误地收录了敏感语料却可能给聊天机器人带来灾难。

此外，聊天机器人往往需要与用户进行一对一的深入交流，聊天机器人的使用者也覆盖了老年人、中年人、青年人、少年儿童等。一些网络用语或许在公开网络上很普遍，但是在与青少年交流时就会变成不良的示范；一些话题或许年轻人能够接受，老年人却可能会完全拒绝。因此，敏感内容的定义也需要变得更加宽泛。

1.2　敏感内容识别方法

敏感内容的识别方法可以按照多种标准划分。其中，比较主流的是敏感词表方法和语义过滤方法。对此可做阐释分述如下。

敏感词表方法，就是构建敏感词库，而后从语料中匹配敏感词，如果能够匹配成功，则说明语料为敏感语料。敏感词表方法往往会受到敏感词表过小、新敏感内容出现以及敏感词的变形体等因素的制约，有很多的改进方法。余敦辉等人[7]提出了基于决策树的敏感词变形体识别算法，通过分析字形、读音等信息，构建决策树，并识别敏感词。

语义过滤方法是指综合语义信息进行过滤。刘梅彦等人[5]先采用主题信息过滤，判断模型是否牵涉敏感话题，再进行倾向性过滤，去除态度敏感的内容。吕滨等人[6]根据语义关系，根据语义框架表示不同，将文本分成了 4 种模式。接下来，分别把已过滤的文本内容和被过滤的文本内容填充语义框架，并计算相似度，从而判断是否需要过滤。

上述方法中，敏感词表方法即使解决了变形体问题，词表的覆盖面以及新敏感词的纳入仍然高度依赖人工操作；语义过滤方法需要使用语义分析工具进行处理，存在误差累积的问题，而且语义框架也是高度依赖人工定义的内容。在聊天机器人的应用背景下，有较大的局限性。

2　敏感内容的定义、分类与分布

2.1　敏感内容的定义与概念

经典看法认为，敏感内容分为主题上的敏感内容，与态度倾向上的敏感内容[5-6]。对于聊天机器人而言，分析敏感内容不能够脱离其依存的客观条件。聊天机器人是一种能够在开放或封闭平台中与用户交互的程序，因此聊天机器人也要“遵纪守法”，不能发表违法、违规或不道德的内容。聊天机器人的设计目的是与用户继续进行持续、愉快的交流，因此聊天机器人也不应该主动发表令用户感到不适的内容，更不能够对用户进行言语上的攻击。最后，聊天机器人往往会面对广大的用户群体，对于一些机构、人物或事物的不恰当评价也容易引起较大的争议乃至商业纠纷，因此也应该尽量避免负面的评价。

通过上述分析，可以发现敏感内容有 3 种层次：首先显著违反法律或道德、不为社会所容忍的内容；其次是在交谈过程中容易让交谈对象感到不舒适、不愉快的内容；最后则是容易引起争议的评价内容。

聊天机器人的主要回复方式分为规则式、检索式与生成式。其中，规则式方法需要人工编写，因此容易控制语料质量，但是无论是检索式、还是生成式聊天机器人，都需要规模较大的语料库，而这样的语料库往往是从网络中爬取构建的。尽管各大网络社区都有尽量避免不友善内容的相关制度与规定，然而，词汇的丰富性、语言表达方式的多样性以及社会热点的实时性使得公开网络上大量存在着敏感内容。

因此，在语料库构建阶段就清洗掉敏感内容，是聊天机器人技术应用的重要步骤。

2.2　敏感内容的分类与分布

敏感内容的分类有 2 个视角。其一是内容的视

角,关注敏感语料具体而言包含什么内容;其二是明显度的视角,关注敏感语料有多容易识别。本文从新浪微博中随机爬取了 360 000 条微博及其下的评论,从中随机抽出了 10 000 对问答。通过人工初步标注,发现敏感内容占比约为 29%。随后,本文又抽取敏感内容中的 500 对问答,分别从上述两个视角考察了敏感内容的分布。

2.2.1 敏感程度角度的分类与分布

从明显程度上说,根据有无敏感词可以进行初步划分;对于前者,又可以根据敏感词的明显程度做进一步划分。总体来说可分成 3 类,即:有明显敏感词的内容、只有隐晦敏感词的内容、不包含敏感词的内容。研究可得,敏感内容在明显程度上的分布见表 1。

表 1 敏感内容在明显程度上的分布

Tab. 1 Distribution of sensitive content in significance

明显程度	占比/%
有明显敏感词	29.06
有隐晦敏感词	26.95
无敏感词	44.99

分析表 1 可以发现,尽管带有明显或隐晦敏感词的数量相当可观,也有相当大一部分数量是没有敏感词的。同时,聊天机器人的语料清洗更重视敏感内容的召回率、而非准确率。因此,在这个任务上,敏感词过滤方法不会得到理想的效果。

2.2.2 内容角度的分类与分布

从内容上说,敏感内容主要分为以下情况:

(1)犯罪、违法、违规内容:牵涉国家、社会、政府机关、政治制度、政策法规、政治人物、宗教信仰、恐怖主义等的内容。

(2)淫秽色情内容:描写性行为,性交,性技巧,性犯罪,与性变态有关的暴力、虐待、侮辱行为以及心理感受的内容,色情淫荡形象等的内容。

(3)不友善内容:针对个人、人群、地域与非公务组织机构的攻击性观点或陈述,对人轻蔑、不尊重的内容。

(4)负面评价:对公司企业、各类产品和社会名人等公共领域进行批评、指责的观点、陈述内容。

(5)消极内容:反映不符合主流价值观的思想倾向,倾向社会阴暗面的内容。

进行统计后发现上述内容的分布情况详见表 2。

分析表 2 可知,不友善的部分占了敏感内容的一半以上,居于首位,这是因为互联网上的聊天有相当一部分是以不尊重的态度进行的;这些内容在互联网平台上或许因可以制造流量与热度而得到容忍,但在聊天机器人中则一样是不合适的内容。仅次于其后的是消极内容,这部分内容谈论的是一些社会的负面信息,在公开网络上往往也是正常的讨论,但也同样不宜出现在聊天机器人的语料库中。

表 2 敏感内容在内容类别上的分布

Tab. 2 Distribution of sensitive content in each category %

内容类别	占敏感内容比例	有明显敏感词的数量占敏感内容比例
犯罪、违法、违规内容	11.24	4.63
淫秽色情内容	7.70	3.58
消极内容	15.70	3.16
负面评价	8.27	2.53
不友善内容	57.09	15.16
总计	100	29.05

接下来,若再考察其中有明显敏感词的项目的比例,就会发现,消极内容、负面内容、不友善内容这三项往往都有相当数量是不带有敏感词的。这也决定了敏感词过滤方法不能很好地识别这些内容。同时,即使是在犯罪违法或淫秽色情这两个类别中,有明显敏感词的内容也不到总体的一半。

2.3 小结

通过统计分析,可以发现敏感内容中有很大的比例不包含敏感词;同时,相比起人们熟悉的犯罪违法违规内容或淫秽色情内容,比例更大的却是不友善内容和消极内容,而且其中的很大部分内容也并不包含敏感词。

3 敏感内容的识别方法研究

本文主要使用传统的敏感词表方法和 bert 文本分类模型[8]进行了敏感内容清洗的实验。除了传统的在准确率 P、召回率 R 以及 F - 值等,本文还引入了 2 个新的评价指标:清洗结果可用度(P_normal)与有效信息留存度(R_normal)。其中,清洗结果可用度是指,清洗完毕后的信息中不敏感内容的占比,可以反映清洗完成后的数据有多少可用,而有效信息留存度则是指不敏感内容在清洗完成后剩余的比例,可以反映保留了多少有效信息。

本文从新浪微博中随机爬取了 360 000 条微博及其下的评论,从中随机抽出了 20 000 对问答,分三次先后标注了 5 000 对、5 000 对和 10 000 对数据。其中,第一次标注的数据作为测试集,后续标注的数据作为训练集。

3.1　敏感词表方法

本文首先从网络收集了 8 个敏感词表(总共含约 7 万词)并集成到一个敏感词表中,同时将集成的词表在收集到的微博全集中统计出现次数,去掉没出现过的词,再按频次从高到低,人工辨别词语的可靠性,进行人工的删除、改写或扩增,保留了2 714 个敏感词,得到优化后的词表。在测试集上分别测试了 2 个词表的表现,详见表 3。

表 3　敏感词表在测试集上的表现

Tab. 3　Performance of sensitive words filter on test set

指标	集成词表	优化词表
P	0.815 2	0.734 5
R	0.126 3	0.179 4
F	0.218 8	0.288 4
P_normal	0.784 6	0.793 2
R_normal	0.991 1	0.979 8

显然,无论是哪种词表,准确率虽然相对较高,但是召回率都很低,远远达不到任务所需要的标准。同时还可以发现,优化后的词表虽然准确率有所降低,却在召回率上有显著的提升,在后续的任务中可以起到更好的作用。

3.2　bert 文本分类模型

本文采用了 Google 公开的 bert 预训练模型,该模型在各项自然语言处理任务中都能起到很好的效果。本文借助这一预训练模型构建文本分类器,先后采用了 5 000 对、10 000 对以及两者组合的数据集进行训练,再在训练集上测试,敏感阈值为 0.5,得到结果见表 4。

表 4　bert 模型在测试集上的表现

Tab. 4　Performance of bert-classifier on test set

指标	5 000-集	10 000-集	5 000+10 000 集
P	0.413 2	0.384 3	0.403 3
R	0.662 2	0.760 7	0.780 1
F	0.508 9	0.510 6	0.531 7
P_normal	0.870 6	0.892 8	0.903 4
R_normal	0.631 1	0.620 5	0.640 7

显然,对标注数据的扩增可以有效提升召回率与清洗结果可用度,数据越多模型性能越好。

3.3　两者相结合的方法

本文进一步尝试结合 bert 模型与敏感词表方法。具体来说,对每对输入内容进行 2 次判断。第一次使用 bert 模型辨别是否为敏感内容,第二次用敏感词表辨别是否敏感内容,任意一次判断为敏感内容就算是敏感内容。得到结果见表 5。

表 5　组合方法在测试集上的表现

Tab. 5　Performance of combined method on test set

指标	5 000+10 000 集	5 000+10 000 集结合集成词表	5 000+10 000 集结合优化词表
P	0.403 3	0.405 3	0.404 0
R	0.780 1	0.795 3	0.806 2
F	0.531 7	0.536 9	0.538 2
P_normal	0.903 4	0.909 0	0.912 5
R_normal	0.640 7	0.636 7	0.629 7

因此,加入优化词表可以在 bert 分类模型的基础上进一步提升性能。

4　敏感内容的扩增方法

根据此前的实验,可以发现,采用分类模型的情况下,扩增训练集的大小可以提升模型性能。而参考各模型的清洗结果可用度,可以发现已有结果的清洗可用度都比较高,因此扩增不敏感内容并不困难,难点在于敏感内容的扩增。

对于敏感内容的扩增有 2 个思路。其一,直接通过敏感词的检索,获得扩增的问答对;其二,通过借助微博文本结构化的信息,从微博中扩增问答对。以下将主要从扩增的内容数量和人工评价得到的合格率两方面来考察敏感语料扩增效果。

4.1　敏感词表直接识别法

本文采用此前优化后的敏感词表,逐个识别微博及其回复构成的所有问答对,提取包含敏感词的问答对,分别考虑只包含 1 个关键词和包含 2 个关键词两种情况。得到的结果见表 6。

表 6　只使用敏感词表时的扩增数量与合格率

Tab. 6　Quantity and qualification rate of pure-filter method

指标	数量	合格率/%
包含 1 个敏感词	41 437	42
包含 2 个敏感词	26 880	88

由表 6 可以看到,随着敏感词数量的增加,敏感词表扩增方法的合格率虽然上升,但同时收集到的敏感语料数量迅速下降。通过对具体结果进行分析,还能发现若干个敏感词对应的敏感内容比例迅速提升。说明单纯使用敏感词表过滤难以构建起大量、稳定的敏感词表。

4.2　结合词表的敏感内容挖掘方法

本文采集的每条微博数据以树状结构保存。父节点为微博以及相关信息,同时有一个以上的子结点,为对该微博的评论;子结点也可以有子结点,为对该条评论的评论。树的深度最大为 3。

本文认为,如果同一条微博下,大多数评论都是

围绕着敏感内容,那么一定有一定数量的微博包含着敏感词,且微博整体也都是明显或隐晦的敏感内容。对于敏感词又可以细分为 2 种,一种是语气上的敏感词,另一种是主题上的敏感词。如果微博中有足够比例的评论都包含语气上的敏感词,那么可以相信微博底下大多数都是语气令人不舒服的评论;而如果微博中包含若干个主题上的敏感词,那么可以相信微博是在围绕着敏感的话题展开对话。

基于上述思考,控制 2 个变量筛选微博数据:一是微博中包含语气敏感词的评论比例 b,二是微博中包含的主题敏感词数目 k。改变 k 时,把 b 固定在 0.1;改变 b 时,把 k 固定在 0,得到实验结果见表 7、表 8。

表 7 不同主题敏感词数目下的扩增数量与合格率

Tab. 7 Quantity and qualification rate of advanced method under different number of topic words

指标	数量	合格率/%
$k = 1, b = 0.1$	442 640	78
$k = 2, b = 0.1$	284 044	80
$k = 3, b = 0.1$	195 439	74
$k = 4, b = 0.1$	143 480	76

表 8 不同含语气敏感词评论比例下的扩增数量与合格率

Tab. 8 Quantity and qualification rate of advanced method under different rate of tone words

指标	数量	合格率/%
$b = 0.10$	2 040 638	64
$b = 0.15$	964 249	58
$b = 0.20$	466 592	72
$b = 0.25$	250 904	78
$b = 0.30$	169 563	67

显然,在扩增的绝对数量上,结合微博结构信息可以增加扩增内容的数量,并且也能够保证合格的敏感内容数量保持在较高的水平,显著优于使用敏感词表直接进行扩增。

此外,根据数据可以发现,无论是哪个筛选指标,单纯提高指标并不会一直提升合格率,指标过高时合格率反而会回落。猜测可能是因为词表中的一些敏感词存在相关关系,把指标提升得过高会导致扩增的数据偏向这些敏感内容。

5 结束语

当前,聊天机器人系统仍然非常依赖于语料库。构建语料库的过程中,敏感内容清洗是一个重要的步骤,且面向聊天机器人系统的敏感内容清洗与一般的敏感内容清洗相比,要求要更严格,过滤失败的风险也会更高。

本文通过分析新浪微博中获取的问答语料数据,得到了 2 个结论:敏感内容多数都以隐晦的形式出现;不同类别的敏感内容占比并不均衡,且总体来说都倾向于隐晦形式。本文提出,衡量敏感内容清洗系统除了使用传统的准确率、召回率和 F - 值,还可以考虑清洗内容可用度与有效内容留存度,以衡量清洗后数据的清洁程度以及有用数据的保留程度。

本文提出了一种基于在无标注数据上优化敏感词表的方式,并实现了一个结合敏感词表与分类模型的敏感内容清洗系统。通过实验发现,对敏感词表使用恰当的清洗方法可以提升其性能,扩充分类模型的训练数据也可以提升分类模型的表现,而且结合分类模型与敏感词表可以实现最好的性能。

本文提出了一种在微博结构语料中,借助敏感词表,提取不包含敏感词的敏感内容的方法,并且在内容抽取数量与质量上都超过了直接使用敏感词表抽取的效果。同时也发现,单纯增加主题敏感词的数量要求或语气敏感评论的比例要求并不能一直提升扩增效果。

本文中最好的扩增方法也只有 80% 的合格率。扩增所得内容中包含的 20% 普通数据的构成与性质,以及具体的去除方法,可以作为进一步研究的内容。

本文提出了清洗低质量敏感词表的方法,而扩增敏感词仍然需要人工介入。如何在已有的研究的基础上,持续扩增敏感词表,也是亟待深入研究的重要内容。

更进一步,如果有办法利用敏感词表持续扩增敏感语料,又可以借助敏感语料的内容持续扩增敏感词表,将可以实现敏感数据清洗的良性循环,也是值得研究的内容。

参考文献

[1] 刘挺. 人机对话技术的进展[R]. 深圳:中国计算机学会,2017.
[2] 张伟男,刘挺. 聊天机器人技术的研究进展[J]. 中国人工智能学会通讯,2016(6):17.
[3] 陈昌凤. 让算法回归人类价值观的本质[J]. 新闻与写作,2018,9(1):1.
[4] 董青岭. 人工智能时代的道德风险与机器伦理[J]. 云梦学刊,2018,39(5):39.
[5] 刘梅彦,黄改娟. 面向信息内容安全的文本过滤模型研究[J]. 中文信息学报,2017,31(2):126.
[6] 吕滨,雷国华,于燕飞,等. 基于语义分析的网络不良信息过滤系统研究[J]. 计算机应用与软件,2010,27(2):283.
[7] 余敦辉,张笑笑,付聪,等. 基于决策树的敏感词变形体识别算法研究及应用[J/OL]. 计算机应用研究:1-7[2019-03-14]. https://doi.org/10.19734/j.issn.1001-3695.2018.11.0792.
[8] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Advances in Neural Information Processing Systems. Long Beach, California, USA: Neural Information Processing Systems Foundation, Inc., 2017:5998.