虚假评论检测研究综述

李璐旸 秦 兵 刘 挺

(哈尔滨工业大学计算机科学与技术学院社会计算与信息检索研究中心 哈尔滨 150001)

摘 要 随着电子商务网站及点评网站的发展,评论信息日益影响着人们的生活.越来越多的网络用户通过发布评论分享消费体验、评价产品的质量,并在做出消费决策时参考其他用户的评论.人们对评论信息的依赖催化了虚假评论的不断涌现.虚假评论,指一些用户出于商业或其他不良动机,在评论中编造不实消费经历、对评价对象的质量等进行鼓吹或诽谤.虚假评论容易对用户的观点或决策产生误导,干扰人们的日常生活.由于人类识别虚假评论的准确率较低,综合运用自然语言处理技术有效检测虚假评论、帮助用户获取真实评论信息,在学术研究及产业应用层面均具有深远意义.对虚假评论检测任务,研究者们主要从虚假评论文本、虚假评论发布者及虚假评论群组三个角度开展研究.该文将依次对三类研究进行归纳分析,具体分别从特征设计、模型方法、数据集、评级指标等方面进行了对比总结.最后对未来研究方向进行了探讨和展望.

关键词 虚假评论检测;虚假评论者检测;合谋欺诈检测;观点挖掘;内容挖掘

中图法分类号 TP393 **DOI 号** 10.11897/SP.J.1016.2018.00946

Survey on Fake Review Detection Research

LI Lu-Yang QIN Bing LIU Ting

(*Research Center for Social Computing and Information Retrieval*,
School of Computer Science and Technology, *Harbin Institute of Technology*, *Harbin* 150001)

Abstract With the development of e-commerce sites and review sites, review information increasingly affects the daily life of people. More and more network users like to post reviews to share the consumption experience and discuss the quality of products; meanwhile, they rely on the reviews from former consumers before making a consumption decision. The dependence of reviews promotes the constant emergence of opinion spam. By the explosive growth of user-generated content, the number of opinion spam in the reviews increases continuously. This phenomenon attracts researchers' attention. Opinion spam is quite different and more crafty than web spam or email spam, which contains opinions of users about products and services. Opinion spam is firstly investigated by Liu et al. who also summarize the opinion spam into different types. In terms of different damages to users, we can further conclude the opinion spam into two types which are deceptive opinion spam (fake review) and product-irrelevant spam. In the former spam, the spammers give undeserving positive reviews or unjust negative reviews to the object for misleading costumers. The latter spam contains no comments about the object. Obviously, the deceptive opinion spam (fake review) is more difficult to detect. The fake reviews are the reviews with untruthful consumption experience and evaluation of products, which may be good reviews about the products of cooperators or bad reviews about the products of competitors driven by commercial profits. The fake reviews likely mislead users on doing decisions which disturb people's daily

收稿日期:2016-09-28;在线出版日期:2017-06-05.本课题得到国家"八六三"前沿技术研究项目(2015AA015407)、国家自然科学基金(61632011,61370164)资助.**李璐旸**,女,1985 年生,博士,中国计算机学会(CCF)会员,主要研究方向为垃圾评论检测、矛盾信息检测、信息可信度.E-mail:lyli@ir.hit.edu.cn.**秦 兵**,女,1968 年生,博士,教授,中国计算机学会(CCF)会员,主要研究领域为文本挖掘、文本分析.**刘 挺**,男,1972 年生,博士,教授,中国计算机学会(CCF)会员,主要研究领域为社会计算、自然语言处理.

life. It is very difficult for people to distinguish fake reviews. In the test of Ott et al., the average accuracy of three human judges is only 57.33%. Hence, the research in detecting fake review is necessary and meaningful. Because of the low accuracy of detecting fake reviews by people, it has far-reaching significance in academic research and industry application that uses technology of natural language processing to resolve the task. The reviews are commonly short documents. The objective of the task is to distinguish the document whether is a spam or a truth. The task can be transformed into a 2-category classification problem. The majority of existing approaches follows Ott et al. and employs machine learning algorithms to build the classifiers. Under this direction, most studies focus on designing effective features to enhance the classification performance. Feature engineering is important, however, and it can hardly learn the inherent law of data from a semantic perspective. In view of the good performance of neural network based models in the natural language processing tasks currently, the document-level representation can be learnt by neural network based models, and be used as features of the review. There are three research directions which are fake reviews detection, fake spammer detection and fake spammer group detection. The paper starts from the introduction of three views of fake review research. Specifically, it makes a conclusion about the feature designing and fake review detection models of the current research works and make a comparison among different types of models. Then it introduces the dataset and evaluation indicators and finally looks into the future of the field.

Keywords　fake review detection; spammer detection; collective spammer detection; opinion mining; content mining

1　引　言

在 Web2.0 时代，用户能够自动生成信息，这些信息中的一类即为产品或服务的评论信息. 现实世界中存在着海量评论信息，而且规模仍在不断增长. 截止 2016 年，美国点评网站 Yelp 拥有超过 1.08 亿条评论信息①，评论的年增长量超过 0.18 亿[1]. 评论信息对用户的观点或消费行为具有导向作用. 相关统计数据表明②，约 81%的美国互联网用户在购买产品前会参考产品评论，其中超过 80%的用户认为评论对他们的购买行为产生了影响. 面对巨大的商业利益，网络涌现出大量虚假评论. 虚假评论(fake review)[2]指对产品或服务进行不符合实际的鼓吹或诽谤，从而达到影响用户的观点或消费行为的目的. 这类评论也称为欺诈评论(deceptive review)或失实观点(untruthful opinion). 如何有效识别虚假评论已经成为亟待解决的网络安全问题之一.

虚假评论分布广泛、危害性大、人工识别困难. 虚假评论广泛分布于食宿、旅游等点评网站及电子商务网站中. 虚假评论在 Yelp 网站中约占 14%～20%[3-4]，在 Tripadvisor、Orbitz、Priceline 及 Expedia 等网站约占 2%～6%. 据报道 2015 年大众点评网的诚信团队封禁违规账号 63 万余个，处理涉及 20 条以上的虚假评论的商家 1.9 万余家③. 研究人员组织三位志愿者对 160 条虚假评论进行人工识别，志愿者倾向于将虚假评论误判为真实评论，识别准确率仅为 53.1%～61.9%[5].

虚假评论这一研究问题最早由 Jindal 和 Liu[2] 于 2008 年提出. 虚假评论属于垃圾评论，垃圾评论是指对用户没有实用价值的评论，包括虚假评论和无关评论. 表 1 列举了真实评论、虚假评论及无关评论的实例. 实例 1 为真实评论，与评价对象内容相关且来自用户真实消费体验. 实例 2 为虚假评论，是被商户雇佣的刷单者编造的评论信息④，属于不可信的垃圾评论. 实例 3、4 为无关评论. 其中实例 3 的内容只与品牌相关、不涉及具体产品或服务. 无关评论还包括广告、问题、回答等. 无关评论相对容易识别，其检测方法简单有效[2]；虚假评论迷惑性高，识别难度大.

① https://www.yelp.com/about
② Kelsey group. Online consumer-generated reviews have-signficant impact onoffline purchase behavior. http://www.comscore.com/press/release.asp?press=1928, 2007
③ http://tech.sina.com.cn/i/2016-03-18/doc-ifxqnskh0974535.shtml
④ http://wenku.baidu.com/link?url=9qIcW6W-UvxwoHxm-0M0HyMD66_nZEdBHpKaFUU-oVVBaAs7CZuoLLkxjeq-E9R0SZZJVKwpdzpKSa0IqFMUUzOWSAmwsQVOznhB4-GR459wXu&_from_mod=copy_login

表 1 真实评论、虚假评论及无关评论实例

序号	评论实例	相关性	可信性	评论类型
1	衣服还行,反正一分钱一分货吧,就是发货太慢了	√	√	真实评论
2	发货速度很快,客服态度好,有问必答,买的很开心. 衣服收到了,超乎意料中的厚度,都没有什么色差. 外面摸上去手感超级好,真的是没买选错. 很赞的店家,质量是太好了. 非常满意	√	×	虚假评论
3	一直喜欢 JNBY 的衣服,品牌值得信赖	×	—	无关评论
4	这衣服你们买时多少钱?	×	—	无关评论
5	我家外贸店均为原单,淘宝店名 xxx,欢迎惠顾	×	—	无关评论

对虚假评论检测研究始于对评论文本的虚假性检测. 该研究问题的难点在于如何对文本、用户等方面的因素及相互间的关系进行更有效的特征挖掘或表示学习,从而提高虚假评论文本检测的准确性. 本文在第 2 节对该研究问题相关的方法、数据集及评价标准进行了介绍和探讨.

随着虚假评论现象的发展和研究的深入,研究者们意识到识别虚假评论者能够更有效地检测虚假评论文本. 虚假评论者指评论者在评论过程中编造虚假的用户体验,或发表与产品或服务的真实情况不符的评价. 虚假评论文本在语法、文体等方面具有较高迷惑性,而虚假评论发布者却在用户属性(如被关注人数)及行为(如个体评分与群体评分的差异)等方面与真实评论发布者有较大的差异性. 对虚假评论发布者的检测研究能够进一步推动虚假评论文本检测研究的发展. 虚假评论者检测研究的难点是如何挖掘更有效的特征刻画虚假评论者与真实评论者之间的差异性. 我们在第 3 节介绍虚假评论者的检测方法、数据集及评价标准.

虚假评论发布者为了扩大对评论对象的控制和影响,近些年以团队合作的形式进行合谋欺诈,这种发布虚假评论的群体被称为虚假评论群组. 相较于单一虚假评论者,虚假评论群组具有更大的社会危害性,检测方法也与前者有很大不同. 对虚假评论群组的检测研究是对虚假评论者检测研究的进一步发展,同时促进虚假评论文本检测系统的性能提升. 研究的难点是如何基于群组在评论内容、评论者行为及群组结构等方面的特性对群组的虚假性进行判断. 虚假评论群组的检测方法及相关数据集、评价标准将在第 4 节中进行介绍.

虚假评论检测的三类研究问题如图 1 所示,与三个检测对象相对应,依次为虚假评论文本检测、虚假评论者检测以及虚假评论群组检测. 评论者的相关信息包括评论文本及评论行为等;评论群组的相关信息包括评论者及网络结构等. 对虚假评论文本的检测研究包括基于文本、基于文本与评论者相结合两个角度的检测研究. 对虚假评论者的检测研究包括基于评论者、基于文本与评论者相结合两个角度的检测研究. 对虚假评论群组的检测研究包括基于群组内容行为分析、基于群组结构分析、基于群组内容行为与结构相结合三个角度的检测研究.

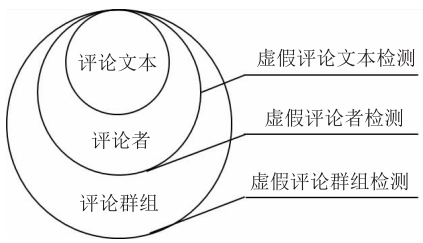

图 1 虚假评论检测研究层次关系图

本文将依据虚假评论的三个研究问题对现有研究工作进行分析总结,具体包括对评论文本、评论者及评论群组的特征分析和模型方法总结,并对方法性能进行了比较. 本文对已有数据资源进行介绍,并对数据集构建方法进行了归纳和分析. 文末对未来的研究方向进行了探讨和展望,为进一步的研究做参考.

2 虚假评论文本检测

虚假评论文本属于评论者编造的一个谎言,其内容建立在虚拟的、不真实的经历或观点上. 尽管谎言编造者会在内容上尽可能模拟现实,在一些语言及行为的细节上仍然会暴露出破绽. 虚假评论文本检测,属于数据挖掘中的内容挖掘(content mining)[6],研究侧重于设计有效的特征和对特征的组合运用. 研究者们依据研究场景的不同,分别从文本、文本与评论者相结合两个角度对虚假评论文本进行检测方法研究,以下将依次进行介绍.

2.1 基于文本分析的虚假评论文本检测

基于文本分析的检测研究包含三类检测方法,分别是基于语法分析、基于语义分析和基于文体元

数据分析的虚假评论文本检测. 下文将依次对有代表性的方法进行总结.

2.1.1　基于语法分析的虚假评论文本检测

语法分析包括对文本进行词袋特征分析及词性特征分析. 一些研究者对虚假评论在语法上的特点进行了研究，并从心理学角度[7]进行解释分析. 我们将在以下语法分析的总结中，对其在语言学或心理学上的相关理论进行简要介绍.

(1) 词袋特征. 评论文本中的单词或连续的多个单词表示文本的词袋特征. 词袋特征又称为 n-gram 特征，unigram、bigram 和 trigram 为常用的词袋特征[2,5,7-9]. 对单词或词组的词频进行统计也用于表示词袋特征[8,10]. 词袋特征对观点挖掘、情感分析等研究方向是一个十分有效的特征. 在虚假评论检测方向，其有效性高于其它文本特征，但在不同数据集中检测效果有明显差异. 在众包平台构造的数据集上，词袋特征能到达到 89.6% 的准确率[5]；而在点评网站的评论数据集上，仅取得 67.8% 的准确率[11]. 点评网站的虚假评论在语言词汇等方面刻意模仿真实评论，词袋特征的单独运用对虚假评论的识别能力不强，与用户行为特征相结合能获得较高识别准确率.

(2) 词性特征. 通过对文本进行词性标注及出现频率统计，提取词性特征. Li 等人[7]分析发现，真实评论和众包方式构造的虚假评论在词性特征上呈现如下特点，前者包含更多的名词、形容词、介词、限定词和连词；而后者包含更多的动词、副词、代词和前位限定词. 这与早期对真实写作(truthful writing)和想象写作(imaginative writing)的分析发现[12-14]基本一致. 然而领域专家编造的虚假评论不满足这一规律，其中包含的名词、形容词和限定词比真实评论多，而动词、副词比真实评论少. 这是由于领域专家编造评论时，模仿真实评论的目的性更强，从产品信息描述、消费体验等细节处模仿真实评论的特点，具有更强的迷惑性. 词性在不同领域评论文本中的分布存在一定的差别性，这与计算语言学的研究结论相一致，即文本中词性的分布与文本类型有关[14]. 然而应用于虚假评论的领域迁移问题，其鲁棒性仍好于词袋特征[7].

研究者基于语法分析获得的特征，运用分类模型如支持向量机及神经网络模型均获得了较好的检测效果. Mukherjee 等人[11]在 Yelp 数据集上采用词袋特征及词性特征，运用支持向量机分类器，在酒店及饭店领域数据集上获得 65.6% 和 67.8% 的准确度. Ott 等人[5,8]、Shojaee 等人[15]及 Li 等人[7]在基于众包平台构造的数据集上，利用词袋特征、词性特征及文体特征等，运用支持向量机分类器获得 84%～89.6% 的检测准确率. 支持向量机将特征向量映射到更高维度空间，从而建立最大间隔超平面，使不同类别的数据点间隔最大、分类误差最小[16]. 支持向量机分类器在解决小样本、非线性及高维特征的问题上表现突出. 景亚鹏[17]在众包方式构造的酒店评论数据集[5,8]上，首先运用信息增益对词袋特征进行特征选择，进而分别通过普通神经网络、DBN-DNN 网络、LBP 网络等三种神经网络模型进行了虚假评论检测. 神经网络模型由大量神经元相互连接形成复杂网络系统，具有自适应和自学习能力，适合于处理数据内在特性不明确的问题.

基于语法分析的虚假评论文本检测模型，通过词袋及词性特征对评论文本的虚假性进行判断，属于该任务的早期研究方法. 该类方法所运用的分类模型中，当有标注数据的规模较小时，支持向量机分类器相较于其它分类器有更突出的检测性能.

2.1.2　基于语义分析的虚假评论文本检测

语义分析运用特征分析方法或语义表示方法对文本在语义层面的信息进行特征提取或抽象表示. 对评论文本进行语义分析，能够获取关于情感词分布的特征及语义表示特征. 应用语义分析的虚假评论文本检测方法有稀疏相加生成模型、神经网络模型及语义语言模型.

(1) 情感分析是一种语义分析方法，通过对情感词分布进行统计分析，从情感极性角度分析文本的语义. 研究表明虚假评论比真实评论包含更多的情感词[7]，语料中表现为比真实评价更积极或更消极. 从心理学角度分析，虚假评论者的目的是鼓吹或诋毁评论对象，评论中情感词的运用能够显示和加强评论的情感极性. 举例说明，表 1 中的实例 1 和实例 2 的情感词分布差异较大. 实例 1 为真实评论，对评价对象的不同方面(如质量、发货时间等)分别有正面或负面的评价，评论内容较为客观. 实例 2 为虚假评论，对评价对象的不同方面均给予好评，包含大量褒义词，具有较强的情感强度.

稀疏相加生成模型运用在虚假评论文本检测中运用情感极性特征发现反常评论信息. Li 等人[7]认为稀疏相加生成模型能够采集多方面因素(如不同领域、有经验或无经验、积极或消极等)，对评论是否为虚假评论进行预测. 稀疏相加生成模型是一种生成贝叶斯方法[18]，可以看作是主题模型[19]和广义相

加模型[20]相结合的模型方法.该方法将不同影响因素作为主题,通过主题间的联合分布概率表示各影响因素间的关系.具体采用拉普拉斯先验对分布稀疏的主题词进行计算表示.实验证明,在多领域数据中处理领域迁移问题,其效果好于支持向量机.模型的具体细节如下:

虚假评论的影响因素集合 Y 包含情感、领域和信息来源三个因素.一个评论的情感有褒贬两种极性;评论内容涉及的领域为宾馆、饭店或医疗中的一种;评论来源包括专业刷单者(employee)、众包平台的任务参与者(turker)及真实用户评论(customer).

$$Y=\{y_{\text{Sentiment}}\in\{positive,negative\},\ y_{\text{Domain}}\in\{hotel,restaurant,doctor\},\ y_{\text{Source}}\in\{employee,turker,customer\}\} \quad (1)$$

每个词条 w 出现的概率 $P(w|i,j,k)$ 计算公式如下,其中 $m^{(w)}$ 表示词条 w 在样本中出现的频率;i,j,k 表示不同影响因素的取值序号;$\eta_{y_{\text{Sentiment}}}^{(i)(w)}$ 表示词条 w 出现在情感极性为 i 值的文档的频率与总出现频率的 log 值的差值;$high\ order$ 表示三个影响因素之间的相互作用关系计算.

$$P(w|i,j,k)\propto\exp(m^{(w)}+\eta_{y_{\text{Sentiment}}}^{(i)(w)}+\eta_{y_{\text{Domain}}}^{(j)(w)}+\eta_{y_{\text{Source}}}^{(k)(w)}+higher\ order) \quad (2)$$

每个评论文本的文档级别的特征 f 的概率 $P(f|i,j,k)$ 计算如下,$m^{(f)}$ 表示样本中特征 f 出现的频率.

$$P(f|i,j,k)\propto\exp(m^{(f)}+\eta_{y_{\text{Sentiment}}}^{(i)(f)}+\eta_{y_{\text{Domain}}}^{(j)(f)}+\eta_{y_{\text{Source}}}^{(k)(f)}+higher\ order) \quad (3)$$

每个评论文档 d 在三方面因素分别取值为 i、j、k 时的概率 $P(d|i,j,k)$ 如下,模型中假设特征和词条出现的概率相互条件独立.

$$P(d|i,j,k)=\prod_{f\in d}P(f|i,j,k)\prod_{w\in d}P(w|i,j,k) \quad (4)$$

在模型训练中,$\eta_y^{(w)}$ 和 $\eta_y^{(f)}$ 通过最大化后验概率学习得到.y_{Source} 对评论文档 d 的来源进行预测,即是否为虚假评论.预测公式如下,在预测中假设领域知识和情感极性已知,令概率最大的 y'_{Source} 即为评论的来源(刷单者、众包志愿者或真实评论者).

$$y_{\text{Source}}=\arg\max_{y'_{\text{Source}}}P(d|y'_{\text{Source}},y_{\text{Sentiment}},y_{\text{Domain}}) \quad (5)$$

情感极性特征同样能够应用于无监督方法对虚假评论文本检测.Jindal 等人[10]认为同一用户对同一品牌的产品全部给予积极评价或消极评价是一种反常行为,对应的评论是虚假评论的可能性大.研究者据此提出了"单一条件规则"(one-condition rules)和"双重条件规则"(two-condition rules)模型,通过概率预测评论文本的虚假性.

(2)语义表示学习是对文本内容的抽象表示对相近的语义表达的泛化处理.基于语义表示的模型通过学习语义表示,将语义表示作为特征进行虚假评论的类别预测.一些研究[21-22]通过神经网络模型学习评论文本的语义表示.语义特征在领域迁移问题上的鲁棒性好于词袋、词性特征及文体特征等.Lau 等人[23]建立语义语言模型识别语义重复的评论,对虚假评论做出预测.

神经网络模型可用于语义表示学习.Li 等人[24]以词向量作为输入,运用卷积神经网络模型(CNN)对评论文本进行语义表示学习,将文本表示向量作为特征进行分类.Li 等人认为评论文本中不同句子的重要性不同,在文本表示学习过程中考虑句子权重能够更好地表示文本语义.模型实现中运用词汇的信息增益计算句子在文档中的权重,在将句子的向量表示合成文档的向量表示过程中,结合句子权重学习文档级向量表示.在多领域混合的数据集上,卷积神经网络模型的性能好于基于长短时间记忆机制的循环神经网络(LSTM-RNN)模型.在跨领域问题上神经网络模型的鲁棒性好于传统分类模型.

(3)语义相似性计算是语义分析的一种常用手段.Lau 等人[23]认为虚假评论存在相互拷贝,通过识别语义重复的评论对虚假评论进行检测.研究者具体在无监督检测方法上进行了探索,提出了语义语言模型.模型对词汇进行同义词泛化并计算评论间的语义相似度,以一定的相似度阈值筛选内容相似的评论,对不相似的文本标注为真实评论,对相似文本通过三位标注者采用投票机制进行人工标注.实验中语义语言模型方法的 AUC 值为 99.87%,明显优于支持向量机的 55.71%.需要指出,该工作中的语料标注方法有待商榷,互不相似的评论文本中同样存在虚假评论,即该方法所标注的虚假评论是虚假评论类别的一个子集.未能对不相似的评论进行虚假评论检测,是该方法的一个缺陷.

基于语义分析的虚假评论文本检测模型,以神经网络模型为代表,通过语义分析对文本进行抽象表示和特征提取,属于该任务较新的研究工作,具有进一步探索的空间.稀疏相加生成模型适用于处理虚假评论文本检测的领域迁移问题.神经网络模型能够对文本进行语义抽象表示学习,并能够融合多种文本特征进行分类,具有良好的扩展性.语义语言

模型可以应用于大量无标注数据，优势是不依赖有标注数据；缺点是应用的启发式方法的规则较为单一，并且将语义相似的评论文本判定为虚假评论将引入误判样例.

2.1.3　基于文体及元数据分析相结合的虚假评论文本检测

对评论进行文体及元数据分析有助于挖掘评论的语言风格及评论者的撰写习惯.提取此类特征能够从文本内容以外的角度分析评论及相应的评论者.

(1)文体特征.文体特征主要用于描述用户写作风格，包括词汇特征和句法特征[15].词汇特征如大写字母的个数、数字的个数、评论的平均长度、第一人称的个数、短单词(如 if、the)的比例等；句法特征如标点符号的数量，功能词的数量如“a”、“the”和“of”等.LIWC①(Linguistic Inquiry and Word Count)是一个常用的文本分析工具[25-28]，能够提取包括文体特征在内的多种文本特征[29].LIWC 将约 4500 关键字表示为 80 维向量，每一维均具有特定心理学含义，Ott 等人对这些维度分成四个类别[5]：文体类信息包括句子平均长度、错误拼写的比例、是否含有不文明词汇等.心理类信息包括与社交、情感、认知、感知、生理相关的特征，也包含时间、空间特征.个性化类信息包括个人的工作、休闲、金钱、宗教等相关的特征.口语类信息包括高频语气词等.

虚假评论在一些文体特征上呈现的规律与心理学对撒谎者的研究结果不完全一致，如虚假评论比真实评论包含更多第一人称代词[7]，这一规律与心理学传统的研究发现相反.心理学研究表明[13,30-32]，撒谎者希望自己与谎话割裂开或由于缺乏真实经历，会不自觉地避免使用第一人称代词.Li 等人[7]对虚假评论中第一人称代词呈现的相反规律进行原因分析，认为虚假评论发布者在评论中编造更多个人消费经历和体会能够令评论读起来更可信.

(2)元数据特征.评论的元数据为评论除文本内容之外的属性特征，如发表日期、时间、评论评级、评论 ID、产品 ID 和反馈信息等.评论的元数据特征有助于识别虚假评论，Li[33]、Hammad[34]及 Mukherjee[11]等研究者均在研究中运用了元数据特征，对评论文本进行数据分析及虚假评论的检测.元数据特征与前述特征相结合能够有效提升虚假评论的识别准确率[11].

现有研究方法一般将文体、元数据分析与语法、语义分析相结合，从而有效提升检测性能.一些检测方法在有标注数据集上，运用经典分类模型如支持向量机、朴素贝叶斯等预测虚假评论文本.还有一些研究面对小规模有标注数据及大量无标注数据，运用半监督方法综合包括文体及元数据在内的多种评论文本特征检测虚假评论.

①Ott 和 Li 等人[5,7]均在虚假评论检测中运用 LIWC 提取文体特征，与词袋特征相结合，检测效果好于单独使用词袋特征.Jindal 和 Liu 等人[2,10]采用逻辑回归模型在亚马逊数据集上采用文体、元数据等特征，并最终结合语法特征获得了 63%～78%的 *AUC* 值，并证明逻辑回归模型在此数据集上的分类性能好于支持向量机和朴素贝叶斯方法.

②一些研究融合多种特征基于半监督方法对虚假评论进行检测[33,35-38].这类方法能够在一定程度上解决虚假评论语料规模较小的问题，方法包括双视图模型、PU 学习模型等.双视图模型(Two-view model)中视图代表不同类型的特征集合，该方法利用多视图的“相容、互补性”对未标注样本进行类别预测和训练集扩充，是一种协同训练算法[39].PU 学习算法(Positive Unlabeled learning)，从少量正例及未标注样本中学习分类器[40-41].算法初始将所有未标注样本作为反例，训练获得的分类器对未标注样本重新进行类别预测，被预测为正例的样本从未标注样本中移除，剩余未标注样本仍作为反例参与下一轮训练.研究者们验证了 PU 学习方法对众多文本分类问题的有效性[42-45].

Li 等人[33]采用双视图方法，综合运用语法特征、语义特征、文体特征及元数据特征等检测虚假评论.视图 1 检测评论文本是否为虚假评论；视图 2 通过检测虚假评论者，进而对其所发布的评论进行虚假评论的预测.基于少量有标注样本，采用两类不同特征分别训练两个分类器，对未标注样本的类别进行预测和标注.扩充有标注数据集后，综合两类特征训练最终的分类器，并对测试集数据进行类别预测.实验证明，通过双视图模型方法标注数据、扩充训练集，其学习的分类器性能好于未扩充的小数据集训练得到的分类器.

Fusilier 等人[36]、Ren 等人[37]和 Li 等人[38]分别提出了基于 PU 学习算法的新模型，获得了超过当时 State-Of-The-Art 方法的分类效果.Fusilier 等人在初始为 100 正例的条件下，训练朴素贝叶斯分类器，最终获得 83.7%的 *F*1 值.Ren 等人提出了 MPIPUL(Mixing Population and Individual property

① http://liwc.wpengine.com/

PU learning)方法,重点识别未标注样本中的间谍样本,即类别特性不显著、易被误分类的样本[46]. MPIPUL 方法首先运用 DPMM[47]对间谍样本进行聚类,继而综合种群性和个体性两种策略对间谍样本类别进行预判和类别标注,最终运用标注数据训练学习支持向量机分类器. Li 等人[38]提出了 CPU 方法(Collective Positive and Unlabeled learning)对 PU-learning 进行改进,将预测结果中的正例加入已有正例集合中,从而在识别反例同时增加正例,更好地训练分类器. 方法通过建立"用户-IP-评论"图,将来自相同 IP 的同一用户与所发布的评论建立关联.

基于文体和元数据分析相结合的虚假评论文本检测模型,融合多种文本特征,是目前较为常见的检测方法. 其中融合多种特征的半监督方法能够解决语料受限的问题. 双视图方法及 PU 学习算法运用少量标注数据训练分类器,并将置信度高的分类结果补充进有标准数据集中,扩充训练语料、提升了分类器的分类能力.

2.2 基于评论文本与评论者相结合的虚假评论文本检测

(1)内容相似性. 对于虚假评论发布者而言,每次编造一条新评论时间成本高,拷贝近似产品的原有评论时间成本低. 虚假评论间的相互拷贝是虚假评论发布者的一个常用手段. Mukherjee 等人在 Yelp 数据集上的分析发现[11],70%以上的虚假评论发布者所发布的评论间相似度大于 0.3,而真实评论发布者所发布的评论间相似度低于 0.18. 对于同一评论者发布的评论进行内容相似度计算,能够反映评论者的评论行为特点. Mukherjee 等人采用最大相似度公式获取内容相似度(Content Similarity, CS). Fei 等人[4]采用平均相似度公式计算评论内容相似度.

$$f_{cs}=\max_{r_i,r_j\in R_a,i<j}\cos(r_i,r_j) \tag{6}$$

(2)评论平均长度. 在 Yelp 数据上的分析发现[11],发布评论的平均长度为小于 135 的虚假评论发布者约占 80%;而 92%的非虚假评论发布者的评论平均长度大于 200. 从心理学角度分析,人们不愿意花费太多的时间在非自发行为上. 虚假评论的长度普遍比真实评论短.

(3)重复评级行为. 对同一产品的多次评级行为,反映了评论者的异常行为[11]. 真实评论者购买同一产品的概率较低,对应的重复评级行为发生的概率也较低.

(4)极端评级行为. 对产品评级最高或最低分数是一种极端评级行为,存在评论者故意褒扬或诋毁产品的可能性. 具体模型设置中,在五星评级体系中对产品评定为一星和五星,是一种极端评级行为[11].

(5)积极评级的比例. 研究表明虚假评论发布者发布的大多数评论均为积极情感的评论. 情感极性通过评级体现,如五星评分制的系统中积极情感对应的产品评级为四星和五星. Mukherjee 等人在 Yelp 数据集上的分析发现[11],85%的虚假评论者发布的评论中 80%的评论为积极评论,而在真实评论者的评论中情感极性的分布更均衡.

(6)评级偏差. 评论者对产品的评级大于该产品的评级均值,则反映了该用户的异常评论行为. 评级偏差(Rating Deviation,RD)为评论者对某产品 p 的评级 v_{rp} 与该产品的平均评级 $\bar{v}_{rp}$ 之间的偏差[11]. Fei 等人[4]对该特征定义的计算公式如下,其中分母的数值为该评级系统最大可能偏差,若评论数据来源的评级系统为五星,则最大的评级偏差为 4.

$$f_{RD}=avg_{p\in P_a}\frac{|v_{rp}-\bar{v}_{rp}|}{rp} \tag{7}$$

(7)重复评论行为. 用户对同一产品提交多条重复的评论,被认为是一种异常行为[4]. 这项特征的设定是否合理有待商榷,Liu 等人[2]指出由于网络连接出现问题或操作失误等原因,同一用户对同一产品提交多次相同的评论不应被认定为发布虚假评论. 识别同一用户的多个 userID[48]有助于检测虚假评论者.

(8)亚马逊确认购买比例. 该特征指某评论者发布的所有评论中有"亚马逊确认购买"标记所占的比例[51]. 亚马逊网站会对有购物记录的评论标注"亚马逊确认购买"标记,有购物记录的评论者对所购物品的评论可靠性高于其他评论.

(9)最大日发布评论数目. 一天内提交大量的评论是一种反常行为[11]. 图 2 所示用户在一天中多次发布评论,属于一种反常行为. 通过统计用户最大日发布评论数(Maximum Number of Reviews, MNR),能够获取虚假评论发布者的异常行为特征. 其中 $MaxRev(a)$ 表示用户最大评论数. Mukherjee 等人在 Yelp 数据集上分析发现,约 75%的虚假评论发布者每天发布 6 条以上的评论,而 90%以上的真实评论者每天的评论数不多于 3 条.

图 2　大众点评网上的评论及评论者信息(左图为某饭店的部分用户评论，左图虚线框圈出的评论为可疑的虚假评论. 右图为此评论对应评论者的个人信息页面，从上至下、从左至右的虚线框依次圈出的为用户头像、用户昵称、用户行为信息以及评论发表时间. 该用户头像、昵称均采用系统默认设置，粉丝数为零，发表评论数量少且发布时间集中在同一天. 综合上述特征，该用户是虚假评论者的可能性很大)

$$f_{MNR}=\frac{MaxRev(a)}{\max\limits_{a\in A}(MaxRev(a))} \tag{8}$$

(10) 突发性发布行为. 研究表明[49-50]虚假评论发布者通常注册时间都较短，发布评论具有突发性. Mukherjee 等人[11]通过定义活跃性窗口，提取突发性发布特征(Reviewing Burstiness，RB). 当最近发布时间 $L(a)$ 与初始发布时间 $F(a)$ 的时间跨度小于一定阈值 τ(28 天[49])，评论者的发布行为属于反常行为.

$$f_{RB}=\begin{cases}0, & L(a)-F(a)>\tau\\ 1-\dfrac{L(a)-F(a)}{\tau}, & \text{其他}\end{cases} \tag{9}$$

(11) 突发性评论比例. 如果一个评论者的突发性评论数占评论者所有评论的比例较大，则该评论者的发布行为属于反常行为. 突发性评论比例(Burst Review Ratio，BRR)，计算突发性评论集 B_a 在评论者所有评论集合 R_a 中所占的比例[51].

$$f_{BRR}=\frac{|B_a|}{|R_a|} \tag{10}$$

(12) 头条评论比例. 由于消费者发布得越早，对其他消费者的影响力越大[51]. 头条评论比例(Ratio of First Reviews，RFR)采集一个用户所有发布的评论中作为对应产品第一个评论者的比例.

$$f_{RFR}=\frac{|\{r\in R_a: r\text{ 为头条评论}\}|}{|R_a|} \tag{11}$$

(13) 早期评论特征. 用户发布的评论是否属于早期评论(Early Time Frame，ETF)[51]，在一定程度上反映了用户的行为特点. 为了最大化地对消费者产生消费导向，虚假评论发布通常都是产品的早期评论发布者.

2.2.1　基于评论文本与评论者相结合的检测方法

在对虚假评论文本的检测研究中，现有研究对文本及评论者的综合分析和利用，主要通过特征级别的融合实现. Mukherjee 等人[11]在 Yelp 数据集上采用支持向量机分类器，运用评论文本特征获得 65.6%～67.8%的准确度. 在加入评论者的特征之后，检测准确度提升至 84.8%～86.1%. 该研究说明评论者特征有助于提升对虚假评论文本的检测能力. Li 等人[33]运用朴素贝叶斯方法和联合训练机制，采用文本及评论者特征对来自点评网站 Epinions 的虚假评论进行检测，获得 61.3%的 $F1$ 值. Hammad 等人[34]在阿拉伯语上运用朴素贝叶斯方法及文本、用户行为特征对 tripadvisor.com、booking.com 和 agoda.ae 等网站的评论数据进行虚假评论检测，获得 99.59%的 $F1$ 值.

Xie 等人[52]运用多时间尺度检测方法，通过时

序分析发现虚假评论集中发布的时间窗口,并认为该类时间窗口内的单评论(singleton review)是虚假评论的可能性很大.其中单评论指该评论的评论者只发布过这一条评论.该研究中运用的检测方法为无监督学习方法,主要利用了评论发布时间及评论者的历史发布记录等特征.Wang 等人[22]对通过张量分解方法学习与评论相关的产品及评论者的向量表示,与词袋特征 bigram 相结合,运用 SVM 对评论的虚假性进行判断.方法对评论相关的所有评论者及产品的相互间关系特征通过矩阵表示,进而运用张量分解技术将每个用户及产品均转化成一个对应的向量表示.这种方法的优势是对全局特征进行向量化表示,有效提升了检测性能.目前在 Yelp 数据集上该方法检测性能最优.

2.3 虚假评论文本数据集

数据集包含真实评论和虚假评论两个类别.目前已有的实验数据集,其真实评论均来自于点评网站.虚假评论根据不同来源可以分为两类,来自真实世界的网站评论和众包平台构造的评论.以下将分别对两类数据资源从标注方法、数据特点等进行对比分析.

2.3.1 网站评论数据集

网站评论数据如表 2 中数据集 1～8 所示.这类资源有如下特点:数据规模大,语法特点及分布能够代表现实世界数据分布,但数据标注存在误差.数据集来自网站,包含评论相对应的评论者信息.应用此类数据集,能够通过用户行为分析识别虚假评论者.由于虚假评论难于识别,缺少有效的识别准则或机制,对虚假评论的有标注数据集的获取是一个难题.现有的网站评论的有标注数据集,其标注方法包括基于规则算法的数据标注及基于人工的数据标注.以下依据不同标注方法对标注方法及数据集进行介绍.

表 2 虚假评论数据资源统计

序号	数据集	数据来源	语言	采集时间	数据量	包含领域	标注数据获取方式
1	580 万 Amazon 评论[2,10,55]	Amazon.com	英语	2006.6	580 万评论 214 万用户 670 万产品	图书、音乐、DVD、工业制品	规则标注
		获取途径 https://www.cs.uic.edu/~liub/FBS/fake-reviews.html					
2	400 万 Amazon 图书评论[56]	Amazon.com	英语	2009	400 万评论 67 万产品	图书	无标注
3	98 万 Amazon 评论[57]	Amazon.com	英语	2013	98 万评论 5 万用户 11 万产品	工业制品	无标注
4	TripAdvisor 宾馆评论[58]	TripAdvisor	英语	2010	3 万评论	宾馆	无标注
5	Epinions 评论[9]	Epinions.com	英语	2011	6 万评论		人工判断
6	Yelp 评论[11]	Yelp.com	英语	2013	6.4 万评论	宾馆、饭店	Yelp 网站过滤算法
7	阿拉伯语评论[34]	tripadvisor.com.eg, booking.com, agoda.ae	阿拉伯语	2007.6～2012.10	2848 评论	宾馆	规则标注
8	大众点评网评论[38]	Dianping.com	汉语	2011.11.1～2013.11.28	9765 评论 9067 用户 5535 个 IP	饭店	大众点评网站过滤算法
9	Ott 等人[5,8]通过众包获取虚假评论	Amazon Mechanical Turk	英语	2011 2013	1600 评论	宾馆	众包
		获取途径 http://myleott.com/op_spam/					
10	Li 等人通过众包获取虚假评论[7]	Amazon Mechanical Turk	英语	2014	3000 评论	宾馆、饭店、医生	众包
		获取途径 http://web.stanford.edu/~jiweil/Code.html					

(1)基于规则算法的有标注数据集

基于规则标注方法不依赖人工,标注成本低,易获得大量标注数据,包含一定的噪音.Jindal 和 Liu 通过分析亚马逊网站的评论语料发现,以下 3 种类型的重复或近似评论为虚假评论的概率很大:

① 同一产品、不同用户 ID 的重复评论;

② 不同产品、同一用户 ID 的重复评论;

③ 不同产品、不同用户 ID 的重复评论.

实际标注中,对以上三类重复评论运用 Jaccard 距离[53]计算评论文本的相似度,相似度大于 0.9 的评论文本标注为虚假评论.基于规则的标注方式有待商榷.Gilbert 等人通过对亚马逊网站上 100 万评

论的分析发现，约 10%～15%的评论与该产品稍早的评论相似[54]. 这说明一些真实用户写评论时趋向于参考或直接复制前人的评论.

基于网站自身过滤算法的标注，可靠性较高，然而算法属于商业机密，没有公开数据集. 数据集 6、8 中的有标注虚假评论数据分别通过 Yelp 网站和大众点评网站自身过滤算法标注. Mukherjee 等人[11]在 Yelp 过滤算法标注的数据集上通过实验对 Yelp 过滤算法进行推测，同时运用了词汇特征和用户行为特征. 用户行为特征应通过对网站公开及内部的数据的分析获得，如用户 IP 地址、地理位置信息、网络及会话日志、鼠标操作、点击行为和评论者在网站的社交行为等. 网站自身过滤算法标注的数据集准确率较高，但召回率低[38]，仍有大量虚假评论未被算法识别、标注.

(2) 基于人工标注的有标注数据集

Ott 等人[8]根据研究报告总结所 30 个虚假评论的判断标准①对点评网站的评论信息进行人工标注. 每条评论由两名标注人员标注，当出现分歧时，由第三名标注人员仲裁. 最终在数据集 5 中，6000 条评论中有 1398 条标注为虚假评论. 人工标注虽然依据一定的评判标准，但仍依赖人的主观判断，不可避免地存在一定数量的误判和漏判样例，识别准确率较低[8]，标注结果可靠性不高.

2.3.2　众包平台评论数据集

众包平台评论数据集(如数据集 9、10)包括两部分数据，真实评论及虚假评论. 真实评论为点评网站经过筛选的评论，而虚假评论来自众包平台. 研究者在众包平台发布评论产品的任务，召集参与者有偿参与任务，通过想象编造产品评论信息作为虚假评论.

Ott 等人[5]在亚马逊土耳其机器人平台发布 400 个人工智能任务(400 Human-Intelligence Tasks)，召集参与者对 20 个旅馆进行评论，共收集 400 条虚假评论. 该众包任务中组织者对参与者设定了撰写规范，不接受有如下特性的评论：包含错误的宾馆名称，内容难以理解，信息过短或剽窃等. 此后，Ott 等人[8]及 Li 等人[7]基于这份数据集，从情感极性(积极和消极)、消费领域(宾馆、饭店和医生)等角度运用众包平台对数据集进行了扩充.

众包平台评论数据集花费一定的人工成本，准确率高，数据集规模较小. 其中的虚假评论能够反映虚假评论者的一些语言及心理特征，但数据集中真实评论与虚假评论的数据分布与真实世界中的分布不符，只能从一定程度上对算法的性能进行验证[11].

2.3.3　数据分析

来自点评网站的评论数据与众包平台构建的评论呈现不同的数据特点. 点评网站的数据集由规则或虚假评论过滤算法进行标注，存在一定数量的误判样例. 众包平台构建的有标注数据集由人工构造，不存在误判样例.

来自点评网站的虚假评论与真实评论在语言特点上比较相近，而众包平台的虚假评论与真实评论的语言差异性较大. 来自众包平台构造的数据中，真实评论与虚假评论的比例与真实世界的实际情况不符，不能完全拟合真实世界的数据分布.

在 Yelp 数据集(数据集 6)上运用 SVM 分类器和 n-gram 特征准确率为 67.8%，远低于同样方法在人工构造的数据集 9 上 89.6%的准确率[11]. Mukherjee 等人从心理学角度分析了原因. Yelp 网站的虚假评论者努力让虚假评论像真实评论一样让人信服，在用词、语法上与真实评论更接近. 人工构造的虚假评论来自于在亚马逊土耳其机器人平台，平台通过付费方式召集参与者完成任务. 任务参与者通过想象编造评论，没有刻意模仿其它真实评论，这些编造的虚假评论与网站真实评论之间的语言差异性大.

2.4　评价指标

对虚假评论的检测研究属于文本分类问题. 在文本分类问题中，对模型性能的评价指标，根据样本正负例分布是否均衡可以分为两类.

(1) 对于虚假评论与真实评论分布均衡的虚假评论数据集，常用评价指标有准确率、精确率、召回率和 $F1$ 值.

准确率(Accuracy)，对算法预测正负例能力的综合评价. Ott 等人[5,8]和 Li 等人[7]在人工构建的数据集上运用准确率评价算法性能. 在正负例均衡的数据集上，适用于运用准确率评价算法有效性. 精确率(Precision)，对算法正确预测正例的能力进行评价. 召回率(Recall)，对算法查找正例的能力进行评价. $F1$ 值($F1$-Score)是精确率和召回率之间的调和值，综合评价算法对正例的预测能力. $F1$ 在虚假

① https://consumerist.com/2010/04/14/how-you-spot-fake-online-reviews/

评论检测研究中是一个常用评价指标[5,7-8,33-34].

(2)对于虚假评论与真实评论分布不均衡[59]的虚假评论数据集,常用评价指标有 ROC 曲线、*AUC* 等.

ROC 曲线(Receiver Operating characteristic Curve),用于描述二分类器的分类性能好坏. ROC 曲线的绘制基于两项指标,真正率(true positive rate)和假正率(false positive rate). 真正率表示正例被正确分类的比例,假正率表示反例被误判为正例的比例. 在二分类问题中,分类器的学习和优化的一个目标是找到两项指标间的平衡. ROC 曲线以真正率为横坐标,假正率为纵坐标,同一曲线上任一点均具有相同感受力,不同点之间区别源自阈值(如逻辑回归模型)设定不同. 曲线越远离对角线、靠近二维坐标图左上角说明分类器的分类能力越强.

AUC(Area Under the receiver operating characteristic Curve)指 ROC 曲线下面积,其值域介于 0.5 和 1 之间. 若对随机挑选的一个正样本及一个负样本通过分类器进行打分,正样本分值大于负样本的概率值即为 *AUC*. *AUC* 值越大说明分类器的准确率越高. 相较于 ROC 曲线,*AUC* 是一种更直观的评价指标. 当样本在不同类别分布不均衡时,准确度不能恰当地反映分类器的分类性能,*AUC* 是一个适用的评价指标. 在基于点评网站的评论数据集上的研究工作[2,23,55],由于正反例分布不均衡,多采用 *AUC* 指标评价方法性能.

2.5 虚假评论文本检测研究小结

基于文本的虚假评论文本检测研究包括从语法、语义及文体元数据三个角度分析文本检测工作. 基于语法分析的虚假评论文本检测方法,运用语法特征通过判别模型如支持向量机、神经网络模型等对评论的虚假性进行预测. 基于语义分析的虚假评论文本检测方法,通过语义特征抽取或语义表示学习方法从文本的语义层面对评论的虚假性进行判断. 基于文体和元数据分析相结合的虚假评论文本检测方法,通过分析评论文本的文体特征、元数据特征,并与语法、语义特征结合使用,综合判断评论的虚假性.

除了从文本角度研究虚假评论,将文本特征与评论者行为等特征相结合也是一类直观、有效的研究角度. 研究结果证明,此类方法相较于基于文本的检测方法,能够有效提升虚假评论文本的检测结果. 在具体研究及应用中,研究场景的不同决定了对两类研究方法的使用. 当研究场景包含评论文本及评论者两种元素时,运用文本与评论者相结合的检测方法能够更有效地检测虚假评论.

我们对虚假评论文本检测任务在部分公开的数据集上进行性能对比,如表 3 所示. 能够看到方法 2、3 相比较,文本特征加入评论者特征后,检测性能有明显提升. 方法 5 运用了张量分解技术对表示特征的矩阵分解为代表评论者及产品的向量,有助于进一步提升虚假评论文本的检测性能.

表 3 虚假评论文本检测方法在部分公开数据集上的性能对比

序号	方法	特征	580 万 Amazon 评论[2,10,60]	Yelp 评论[11](饭店领域)	Ott 等人[5,8]通过众包获取虚假评论	Li 等人通过众包获取虚假评论[7](饭店领域)
1	逻辑回归[2]	评论文本、评论者及产品特征	*AUC*=78%	—	—	—
2	支持向量机[5]	文本特征(Bigram+LIWC)	—	—	*Accuracy*=89.8%	—
3	支持向量机[11]	文本特征(Bigram)	—	*Accuracy*=68.5%	*Accuracy*=88.8%	—
4	支持向量机[11]	文本特征(Bigram)+评论者特征	—	*Accuracy*=86.1%	—	—
5	支持向量机+张量分解[22]	文本特征+评论者特征+产品特征	—	*Accuracy*=89.9%	—	—
6	支持向量机[7]	文本特征(Unigram)	—	—	—	*Accuracy*=81.7%

3 虚假评论者检测

虚假评论者指发布虚假评论(一条或以上),对评论对象进行恶意诋毁或不符合实际的褒扬的评论者. 虚假评论者检测是近些年的研究热点,已有研究工作可以归纳为两个研究角度,基于评论者的检测研究及基于评论文本与评论者相结合的检测研究.

3.1　基于评论者分析的虚假评论者检测

这类方法通过挖掘评论者特征，分析评论者的反常行为对虚假评论者进行预测. Fei 等人[4]认为爆发性涌现的评论者更可能为虚假评论者，通过时序分析检测虚假评论者. 研究者运用马尔可夫随机场(MRF)模型基于评论者之间的关系构建评论者网络，将评论者作为观察节点，每个评论者的真实类别为隐含节点，对一个爆发期内共同出现的评论者用边相连接、建立关联关系. 运用的特征有亚马逊确认购买比例 f_{RAVP}、评级偏差 f_{RD}、突发性发布行为 f_{RB}、突发性评论比例 f_{BRR} 和内容相似性 f_{cs}. 由于亚马逊确认购买比例这一特征对虚假评论者的检测具有较高的识别能力，将这一特征作为评论的初始状态. 运用如下的公式计算虚假评论指示值(Overall Spamming Indicator，OSI)预测评论者发布虚假评论行为.

$$OSI(a)=\frac{f_{RAVP}+f_{RD}+f_{RB}+f_{BRR}+f_{cs}}{4} \tag{12}$$

3.2　基于评论文本与评论者相结合的检测

现实世界中，虚假评论者的检测涉及评论者、评论文本等多方面因素. 评论者之间[61]、评论者与评论文本及评论对象间均存在网络拓扑结构. 研究者们对评论者、评论文本，包括评论对象构建关系网络，通过基于图的方法对虚假评论者进行检测. 现有方法包括基于 HITS 算法的检测模型和基于马尔可夫随机场的检测模型.

(1) Wang 等人[62-63]构建了“评论者-产品”网络、“评论者-评论”网络及“评论-产品”网络，如图 3 所示. 文中定义了评论者的可靠性(trustiness)、评论的真实性(honesty)以及产品的信誉度(reliability). 这三种变量的取值分别由其他两种变量决定，并运用类似于 HITS 算法[64]进行迭代求解.

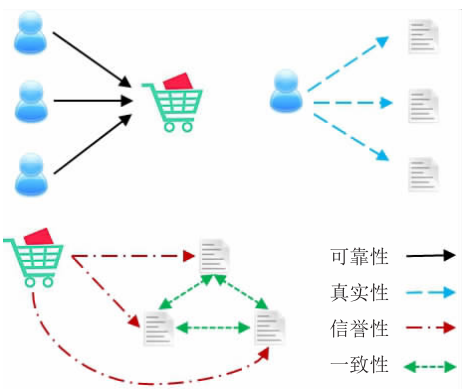

图 3　产品评论图中各类型节点间的关系图

评论者的可靠性与所发评论的真实性正相关，即所发评论真实性越高，评论者可靠性越高. 评论数较少时，评论的真实性对评论者的可靠性影响更大. 假设 $T(i)$ 表示评论者 i 的可靠性，H_i 表示评论者 i 所发出的所有评论的真实性的总和，$T(i)$ 的计算公式如下：

$$T(i)=\frac{2}{1+e^{-H_i}}-1 \tag{13}$$

评论的真实性与两个因素有关，产品的信誉度以及该评论与一定时间段内该产品其它评论的一致性. $H(r)$ 表示评论 r 的真实性，$R(\Gamma_r)$ 表示评论 r 所评论的产品 Γ_r 的信誉度，$S_{r,a}$ 代表评论内容与 r 一致的评论集合，$S_{r,d}$ 代表与评论 r 内容不一致的评论集合. $A(r,\Delta t)$ 通过采集时间窗口大小为 Δt 的评论集合 S_r 计算评论 r 与窗口时间内所有评论的一致性，$A_n(r,\Delta t)$ 为归一化到(−1,1)值域后的评论一致性值.

$$A(r,\Delta t)=\sum_{i\in S_{r,a}}T(i)-\sum_{j\in S_{r,d}}T(j) \tag{14}$$

$$A_n(v,\Delta t)=\frac{2}{1+e^{-A(v,\Delta t)}}-1 \tag{15}$$

$$H(r)=|R(\Gamma_r)|A_n(v,\Delta t) \tag{16}$$

研究者认为产品的信誉度 $R(\Gamma_r)$ 与该产品所有评论者的可靠性相关，具体呈对数关系. 其中 U_S 表示该产品所有评论的集合，Ψ_r 表示评论 r 对应的评级，μ 表示评级的中值(如五星评级系统中的三星).

$$R(\Gamma_r)=\frac{2}{1+e^{-\theta}}-1 \tag{17}$$

$$\theta=\sum_{r\in U_S,T(r)>0}T(r)(\Psi_r-\mu) \tag{18}$$

(2) Rayana 等人[65]运用图结构表示“评论者-评论-产品”网络，结合个体评论者特征如评论发布时间、对产品的评级等，运用马尔可夫随机场模型 SpEagle 对评论者、评论及产品这三类元素的标签类别同时进行预测. 马尔可夫随机场(MRF)模型包括两种类型的节点，观察节点(observed node)和隐含节点(hidden node). 观察节点表示数据能够观察到的值. 每个隐含节点与一个观察节点对应，表示数据隐含的真实状态[66]. 研究者通过此方法预测用户是否为虚假评论者，预测产品是否被虚假评论者攻击，并预测评论是否为虚假评论.

(3) Akoglu 等人[67]对评论者及产品构建二分图，提出了基于马尔可夫随机场的 FraudEagle 模型. 图结构中积极(或消极)评论作为有权重的边将

评论者和产品相连接,进而通过对隐含节点的类别预测对网络中评论者及对应的评论进行欺诈检测.

基于 HITS 算法的检测模型能够通过计算几类因素之间的相关性较好地解决该检测问题.缺点是对新加入的样本需要重新进行全网节点的计算,扩展性不强.基于马尔可夫随机场的检测模型,优势在于运用网络图结构能够更清晰地表示评论、产品和评论者之间的关联关系,同时对于未标注数据有很好的扩展性,能够通过网络结构进行信誉度传递,进而对虚假评论进行识别.缺点是目前基于特征的组合运算较为简单,应综合考虑各个特征的效用及特征之间是否存在非线性关系等因素,提出更有效的特征组合计算方法.

3.3 虚假评论者数据集

3.3.1 亚马逊评论数据

Lim 等人[51]在亚马逊购物网站获得制成品评论数据集(MProducts).数据集包含 11038 位评论者.请三位标注者对其中的 50 个评论者进行了虚假性评论发布者与真实评论发布者两种类别的标注.标注人员基于每个评论者的 10 条评论及行为特征(如对同类产品的重复评论行为、对产品的评级与平均评级的差异、评论内容是否存在冗余性等)进行综合判断,对评论者进行标注,并采用多数投票机制对每个评论者产生最终的类别标签.

3.3.2 Resellerratings 数据集

Wang 等人[62]在 resellerratings.com 网站收集 343603 评论者,通过算法筛选出 100 名疑似虚假评论发布者.为三位标注人员提供有关评论者的三类信息,标注人员基于这些元数据信息对 100 名疑似虚假评论发布者进行标注.三类评论者的元数据信息有:该评论者对商品的评分是否与商品的平均评分相反;评论者对商品的评分是否与 Better Business Bureaus 机构①给出的评分相反;评论者对商品的评分与大众对商品的评价相反.若一个评论者被两名或以上的标注人员标注为虚假评论者,则被归类为虚假评论者.通过人工标注,该数据集包含 49 个虚假评论发布者及 51 个真实评论发布者.

3.3.3 YelpChi、YelpNYC 及 YelpZip 数据集

Yelp 网站通过设计的过滤算法对用户发布的评论进行过滤,在商品的评论区显示的是过滤后的可靠评论.点击商品页面底部的链接,能够看到被过滤掉的疑似虚假评论.研究者认为发布了虚假评论的用户即为虚假评论发布者,发布可靠评论且未发布虚假评论的用户即为真实评论发布者.通过这种方式,Mukherjee 等人[11]收集了 YelpChi 数据集,Rayana 等人[65]收集了 YelpNYC 及 YelpZip 数据集.YelpChi 数据集包含 38036 个虚假评论者,占数据集所有用户的 20.33%;YelpNYC 包含 160225 个虚假评论者,占数据集所有用户的 17.79%;YelpZip 包含 260277 个虚假评论者,占数据集所有用户的 23.91%.

3.4 评价指标

一些研究方法以概率预测的方式对评论者是否为虚假评论发布者进行判断.分类问题转化成排序问题,可能性大的评论者排序靠前,因此研究者基于信息检索任务的评价机制对虚假评论发布者的检测结果进行评价.评价虚假评论者检测方法的性能的标准包括:精确率(Precision)、召回率(Recall)、ROC 曲线、*AUC*、*NDCG@k*(Normalized Discounted Cumulative Gain)等.

$$NDCG@k=\frac{DCG@k}{IDCG@k} \tag{19}$$

$$DCG@k=\sum_{i=1}^{k}\frac{2^{l_i}}{\log_2(i+1)} \tag{20}$$

$NDCG@k$ 表示前 k 个结果与真实结果的拟合程度,值越大说明方法的性能越好[65].其中 l_i 表示排序在第 i 个评论者的类别(1 表示虚假评论者;0 表示真实评论者).$IDCG@k$ 是运用 $DCG@k$ 公式对正确的排序结果进行计算,其中 l_i 均取值为 1.

3.5 虚假评论者检测研究小结

基于评论者的检测模型通过在特定时间间隔内采集爆发式涌现的评论者,在检测中运用时序分析,能够从时间维度上有效区分真实评论者与虚假评论者.该方法若能结合评论的文本特征,检测性能将会得到进一步提升.

基于评论与评论者相结合的检测模型,运用图模型表示评论者、评论及产品信息组成的网络结构,利用节点间可信度的传递性进行虚假评论者的检测.该类方法是对真实世界的近似模拟,是符合直觉、检测能力较好的一类模型.

我们在几个标准数据集上对能进行对比的方法及结果进行了列举,如表 4 所示.能够看到目前在 Yelp 系列数据集上,SpEagle 方法获得最佳的检测结果.

① Better Business Bureaus 是一个对产品及品牌的信誉度进行评分的公司.它收集商业报告,提醒公众一些消费陷阱,加强消费者与企业之间的相互信任.

表 4　虚假评论发布者检测方法在部分公开数据集上的性能对比

序号	方法	特征	Resellerratings 数据集	Yelp'Chi	Yelp'NYC	Yelp'Zip
1	规则	评论者发布冗余评论[51]	*Precision*=3%			
2	FraudEagle[67]	对评论者及产品构成的网络抽取特征	—	*AUC*=61.24%	*AUC*=60.62%	*AUC*=61.75%
3	Wang 等人的迭代算法[63]	产品、评论及评论者两两间关系	*Precision*=49%	*AUC*=61.67%	*AUC*=62.07%	*AUC*=65.54%
4	SpEagle[65]	文本、评论者及相互关系的相关特征	—	*AUC*=69.05%	*AUC*=65.75%	*AUC*=67.10%

4　虚假评论群组检测

虚假评论群组(collusive spammers 或 spammer groups)[49,68]指一定数量的评论者通过协调运作有组织地对某个产品进行虚假评论. 虚假评论群组的社会危害性较大. 这类群组通常在一定时间间隔内，对评价对象批量产生大量不实评价，影响甚至控制评价对象的声誉. 虚假评论群组的行为称为合谋欺诈[66]. 由于虚假评论群组发布评论时间较集中，所发布的虚假评论在同一产品中往往分布在相邻位置、相互掩护，对虚假评论群组的检测难度较大.

虚假评论群组检测研究的对象是群组，研究内容是通过分析群组的结构、发布内容及行为特征判断群组的虚假性. 也就是说虚假评论群组检测任务不包含群组的识别. 群组的识别属于该任务的前续工作，Mukherjee 等研究者在构造该任务实验数据集过程中，采用"共同评论多个产品的评论者为一个群组"机制，识别群组[49].

群组由群组成员以一定的网络结构组成，如图 4 所示，群组成员的评论内容及行为以及群组网路结构特点对群组的虚假性判断均有影响. 对虚假评论群组的检测研究主要包括三个角度的研究，基于群组内容及行为分析的虚假评论群组检测、基于群组结构分析的虚假评论群组检测和基于内容行为分析与网络结构相结合的虚假评论群组检测研究. 我们将依次对特征及模型方法进行介绍，并对方法进行对比分析.

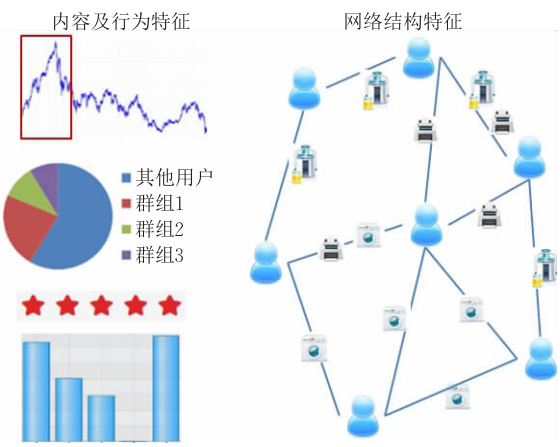

图 4　虚假评论群组影响因素图

4.1　基于群组内容及行为分析的虚假评论群组检测

研究者认为构成虚假评论群组的评论者之间有相近的特性，通过挖掘评论者之间的关联关系或基于评论者的群组特征进行聚类能够对虚假评论者构成的群组进行检测.

4.1.1　群组内容及行为分析

群组的内容及行为分析包括对群组内成员所发布的评论的内容分析，及对群组成员的行为分析. 群组内容特征包括群组内容相似性和群组成员内容相似性. 群组行为特征包括群组时间窗口、群组评级偏差、群组支持率和群组早期评论特征.

(1) 群组内容相似性. 群组内评论者发表虚假评论时，存在相互拷贝的现象. 重复或近似重复的评论有较大的可能来自虚假评论群组[51]. g 表示群组，$c(m_i,p)$ 表示群组中评论者 m_i 对产品 p 的评论内容，采用 cosine 计算一对评论者 m_i 和 m_j 针对产品 p 的内容相似度，$h_{CS}(g,p)$ 计算群组中评论者们关于产品 p 的平均相似度，表示群组在该产品上的内容相似度表现. 群组内容相似性(Group Content Similarity，GCS)在所评论的产品集合 P_g 中取评论内容相似度最大值作为该群组的内容相似度值.

$$f_{GCS}(g)=\max_{p\in P_g}(h_{CS}(g,p)) \tag{21}$$

$$h_{CS}(g,p)=avg_{m_i,m_j\in g,i<j}(\text{cosine}(c(m_i,p),c(m_j,p))) \tag{22}$$

(2) 群组成员内容相似性. 若群组中评论者不相互协同，分别独自发布虚假评论，则倾向于拷贝各自的已发布评论[51]. $h_{CS}(g,m)$ 计算同一评论者 m 对不同产品 p 的平均内容相似度. 群组成员内容相似度(Group Member Content Similarity，GMCS)对群组 g 中所有成员的内容相似度取均值，用于表示群组成员对自身评论拷贝情况的综合表现.

$$f_{GMCS}=\frac{\sum_{m\in g}h_{CS}(g,m)}{|g|} \tag{23}$$

$$h_{CS}(g,m)=avg_{p_i,p_j\in P_g,i<j}(\text{cosine}(c(m,p_i),c(m,p_j))) \tag{24}$$

(3)群组时间窗口. 在虚假评论群组中,评论者们倾向于在一个时间段内对某个评价对象频繁发布评论. 群组时间窗口(Group Time Window,GTW)反映了评论群组 g 的评论行为在一定时间跨度上的表现[51]. $h_{GTW}(g,p)$是该群组对产品 p 的评论时间窗口,其中 $L(g,p)$为群组中最新评论的发布时间,$F(g,p)$为最早的评论的发布时间,τ 为阈值. $f_{GTW}(g)$是群组 g 对所评论的产品集合 P_g 中所有产品的评论时间窗口中最大的窗口值.

$$f_{GTW}(g)=\max_{p\in P_g}(h_{GTW}(g,p)) \tag{25}$$

$$h_{GTW}(g,p)=\begin{cases}0, & \text{若 } L(g,p)-F(g,p)<\tau\\ 1-\dfrac{L(g,p)-F(g,p)}{\tau}, & \text{其他}\end{cases} \tag{26}$$

(4)群组评级偏差. 虚假评论群组对产品的评级会与真实用户的评级存在偏差,偏差越大,群组的合谋欺骗行为越显著[51]. 群组偏差(Group Deviation,GD)$f_{GD}(g)$是群组 g 在所有所评论产品的评论行为中的一个最差表现. $h_{GD}(g,p)$为群组 g 对产品 p 的评级偏差,其中 v_{gp} 表示群组 g 对产品 p 的平均评级,$\bar{v}_{gp}$表示其它评论对产品 p 的平均评级. 若评级系统为五星评级,则最大可能的评级偏差为 4,$h_{GD}(g,p)$通过除以最大可能的评级偏差进行正规化.

$$f_{GD}(g)=\max_{p\in P_g}(h_{GD}(g,p)) \tag{27}$$

$$h_{GD}(g,p)=\frac{|v_{gp}-\bar{v}_{gp}|}{gp} \tag{28}$$

(5)群组支持率. 如前所述,群组的构成是基于评论者行为的一致性,在越多的产品上产生一致性,群组是合谋欺诈的可能性越高. 群组支持率(Group Support Count,GSC)反映了群组 g 在所有群组中归一化后的表现[51].

(6)群组早期评论特征. 在产品对应的评论集合中,出现在早期的虚假评论会对产品的声誉和销售产生更大的影响[51]. 若群组中成员发表了最初的评论,则群组更容易控制产品评论的情感倾向性. $A(p)$表示产品 p 最早评论时间,$L(g,p)$表示群组 g 对产品 p 的最近一次评论时间,β 为表示时间间隔的阈值(Mukherjee 等人[49]设定为 6 个月). $h_{GTF}(g,p)$体现了群组对产品 p 的早期评论特征的表现,值越大说明群组发表的评论时间越早. 若群组的最后发表评论时间与该产品能够发表评论的时间跨度大于阈值,则说明群组发表的评论并非都集中于早期,从一个侧面说明群组进行虚假评论的动机不显著. 群组早期评论特征(Group Early Time Frame,GETF)采用了群组评论的所有产品 P_g 中的早期时间特征最大值.

$$f_{GETF}(g)=\max_{p\in P_g}(h_{GTF}(g,p)) \tag{29}$$

$$h_{GTF}(g,p)=\begin{cases}0, & \text{若 } L(g,p)-A(p)<\beta\\ 1-\dfrac{L(g,p)-A(p)}{\beta}, & \text{其他}\end{cases} \tag{30}$$

4.1.2 基于群组内容及行为分析的检测方法

这类方法从群组成员角度,通过分析群组内容及行为特征,发现群组成员之间的关联性,进而通过排序或聚类方法识别虚假评论群组. 以下对有代表性的方法进行介绍.

(1)研究者[69]认为评论者之间关联性与多个特征有关,包括评论者对共同产品的评级偏差、评论的时间偏差、评论内容的相似度、对品牌的评级偏差、基于品牌的评论时间偏差、评论行为的同质性及评论者账号的使用时间的同质性. Xu 和 Zhang[69]提出了运用多种成对特征(pairwise features)挖掘评论者之间的关系,通过排序方法 FraudInformer 对虚假评论群组进行检测. 研究者认为评论者属于虚假评论者的概率越高,属于虚假评论群组的概率也越高. 两个有虚假评论倾向的评论者若关联性强则更可能构成合谋欺诈行为. 假设 s_i 表示评论者 i 的虚假分值,个体评论者的虚假分值由该评论者的相邻评论者的虚假分值 s_j、评论者之间的关联性 $\Phi(i,j)$及关联性的置信度 $\Omega_j(i,R)$共同决定. 方法首先计算评论者之间关联性及关联性的置信度,进而运用马尔可夫随机游走模型[70]进行全局排序.

$$s_i=\sum_{k\in P_i}\sum_{v_j\in R_k}s_j\cdot\Phi(i,j)\cdot\Omega_j(i,R_k) \tag{31}$$

(2)Xu 等人[68]利用文本特征和评论者的特征计算成对评论者间的相似度,运用改进的 KNN 聚类算法进行聚类,并选择 k 个最相似评论者通过投票机制判断是否为虚假评论群组.

基于评论者的虚假评论群组检测方法中,排序方法采用 FraudInformer,算法思想直观、扩展性好. 缺点是方法需要人工对判定类别的阈值进行设定,对专家知识依赖性较强. 基于聚类的方法将特性相

同的评论者聚为一类，方法较为直观. 缺点是聚类过程中簇的数目的确定是一个难点，预先指定固定数目的簇会限制最终聚类的效果.

4.2 结合群组结构分析的虚假评论群组检测

由于群组具备天然的网络结构特征，虚假评论群组在网络结构上具有一定的特异性. 结合网络结构分析的虚假评论群组检测方法能够有效提升检测性能.

4.2.1 群组结构分析

群组结构特征对群组所构成的网络结构提取特征，包括群组规模、群组规模比例、群组紧密度、相邻节点多样性和自相似性.

(1) 群组规模. 群组规模(Group Size，GS)反映了群组成为虚假评论群组的可能性. 当群组成员间未通过协同合作发布评论时，群组规模越大，群组成员行为一致的概率越小. 反言之，当规模较大的群组出现群组成员一致行为时，群组进行合谋欺诈的概率较高. G 表示所有候选群组的集合，g 表示群组 g.

$$f_{GS}=\frac{|g|}{\max\limits_{g_i\in G}(|g_i|)} \tag{32}$$

(2) 群组规模比例. 群组中评论者的数目在产品所有评论者中所占的比例反映了群组虚假评论行为对产品声誉的控制能力. A_P 表示产品 p 对应的所有评论者，h_{GSR} 是群组 g 在产品 p 的所有评论者中所占的大小比例. 群组规模比例(Group Size Ratio，GSR) $f_{GSR}(g)$ 指群组成员在所评论的所有产品 P_g 中平均所占比例[74].

$$f_{GSR}(g)=avg_{p\in P_g}(h_{GSR}(g,p)) \tag{33}$$

$$h_{GSR}=\frac{|g|}{|A_P|} \tag{34}$$

(3) 群组紧密度. 群组中各个成员之间是否紧密关联反映了群组合谋欺诈的可能性. $T(m,p)$ 表示群组成员 m 对产品 p 的评论时间，$L(g,p)$ 和 $F(g,p)$ 分别表示群组 g 对产品 p 的最近及最早的评论时间. $avg(g,m)$ 反映了群组 g 中所有成员在产品 p 上的平均评论时间跨度. 群组紧密度(Individual Member Coupling in a group，IMC) $f_{IMC}(g,m)$ 反映了群组中成员 m 与群组中其它成员的紧密程度，反映了成员间的行为相似程度[74]. 公式采用了在群组所评论的所有产品上的平均紧密度作为最终取值.

$$f_{IMC}(g,m)=avg_{p\in P_g}\left(\frac{|(T(m,p)-F(g,p))-avg(g,m)|}{L(g,p)-F(g,p)}\right) \tag{35}$$

$$avg(g,m)=\frac{\sum\limits_{m_i\in G-\{m\}}(T(m_i,p)-F(g,p))}{|g|-1} \tag{36}$$

(4) 相邻节点多样性. 在群组的虚假性检测研究中，评论者构成了群组网络中的各个节点. 在真实网络环境中，节点与相邻节点在一些属性或行为上应具有差异性，即节点之间不应过度依赖. 研究者采用熵 H 衡量相邻节点的多样性，$p_k^{(i)}$ 表示节点 i 在产品 k 对应的评论者网络中的离散概率分. 具体通过节点的度及 $pagerank$ 值衡量节点的多样性，运用熵 H_{deg}、H_{pr} 来衡量节点的度和 $pagerank$ 值两个特征.

$$f(H(i))=P(H\leq H(i)) \tag{37}$$

$$H(i)=-\sum_{k=1}^{K}p_k^{(i)}\log p_k^{(i)} \tag{38}$$

(5) 自相似性. 网络的局部与整体之间在一些属性上具有相似性[71-73]. 具体通过节点的度和 $pagerank$ 值作为特征进行衡量，公式中运用 KL 散度(相对熵) KL_{deg}、KL_{pr} 衡量网络中节点的自相似性[74].

$$f(KL(i))=1-P(KL\leq KL(i)) \tag{39}$$

$$KL(P^{(i)}\|Q)=-\sum_{k=1}^{K}p_k^{(i)}\log\frac{p_k^{(i)}}{q_k} \tag{40}$$

4.2.2 结合群组结构分析的检测方法

一类方法通过分析评论者所构成的网络结构，提取群组结构特征，基于聚类方法将具有相似群组结构特征的评论者聚为一类，进而运用启发式或投票机制预测群组是否为虚假评论群组.

Ye 和 Akoglu[74] 提出了基于群组网络结构分析检测方法. 方法基于 2-hop 子图发现评论者作为群组成员的反常行为并运用层次聚类方法 GroupStrainer 检测虚假评论群组. 研究者在评论者构成的网络结构中，运用相邻节点多样性和节点与网络自相似性两类群组特征，通过两类特征对评论者的网络足迹进行分析. 评论者在网络中的足迹(Network Footprint Score，NFS)反映了行为的反常性，具体通过对两类特征计算获得.

$$f(NFS(i))=1-\sqrt{\frac{f(H_{deg}(i))^2+f(H_{pr}(i))^2+f(KL_{deg}(i))^2+f(KL_{pr}(i))^2}{4}} \tag{41}$$

还有一类方法同时对群组成员及群组结构进行分析，即将自底向上和自顶向下的分析相结合. 通过群组成员分析提取群组内容及行为特征，并运用群组结构信息通过图模型检测虚假评论群组.

(1) Mukherjee 等人[49] 运用排序模型 GSRank

进行虚假评论群组检测.方法采用多种评论群组特征,计算候选群组属于虚假评论群组的概率值,并通过排序预测虚假评论群组[75].方法对产品集合 P、群组集合 G 及群组成员集合 M,建立三个二元关系模型,产品-群组模型(Product-Group,PG)、成员-产品模型(Member-Product,MP)和群组-成员(Group-Member,GM)模型.三个模型通过矩阵存储相关性系数,构建了群组、成员及产品之间的关联性.

在计算产品-群组关联性矩阵 $\boldsymbol{W}_{PG}$ 时,运用了群组时间窗口 f_{GTW}、群组偏差 f_{GD}、群组早期评论特征 f_{GETF}、群组成员内容相似度 f_{GMCS} 及群组规模比例 f_{GSR} 五个群组特征(具体公式参见 4.1.1 节及 4.2.1 节).

$$\boldsymbol{W}_{PG}=\left[\frac{1}{5}(f_{GTW}+f_{GD}+f_{GETF}+f_{GMCS}+f_{GSR})\right]_{|P|\times|G|} \tag{42}$$

在计算成员-产品之间的关联性矩阵 $\boldsymbol{W}_{MP}$ 时,运用了评论者特征中的评级偏差 f_{RD}、内容相似性 f_{CS} 及早期评论特征 f_{ETF}.

$$\boldsymbol{W}_{MP}=\left[\frac{1}{3}(f_{RD}+f_{CS}+f_{ETF})\right]_{|M|\times|P|} \tag{43}$$

在计算群组-成员关联性矩阵 $\boldsymbol{W}_{GM}$ 时,运用了群组紧密度 f_{IMC}、群组规模 f_{GS} 及群组支持率 f_{GSC} 三个群组特征.

$$\boldsymbol{W}_{GM}=\left[\frac{1}{3}(f_{IMC}+f_{GS}+f_{GSC})\right]_{|M|\times|P|} \tag{44}$$

向量 $\boldsymbol{V}_G$ 表示群组集合 G 中各群组是虚假评论群组的概率,$\boldsymbol{Z}$ 表示三类关系模型之间的关联关系.该方法通过 GSRank 模型反复迭代求解 $\boldsymbol{V}_G$.

$$V_G^{(t)}=(\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z})V_G^{(t-1)} \tag{45}$$

$$\boldsymbol{Z}=\boldsymbol{W}_{GM}\boldsymbol{W}_{MP}\boldsymbol{W}_{PG} \tag{46}$$

(2) Xu 等人[68]基于双马尔可夫网络[76]提出了欺诈图模型(Colluder Graph Model,CG)对虚假评论群组进行检测研究.方法将评论者作为节点构成网络,关联性强的评论者互为相邻节点,同时对用户-属性建立对应关系,将问题转化为运用概率图模型预测虚假评论群组.方法基于的假设是,通过评论者发布的评论文本及网络中相邻评论者的属性能够预测网络中的评论者节点的类别.该方法将每个节点的真实类别作为待预测的隐含变量 $L=\{L_i\}_{i=1}^{m}$,$\mathcal{E}$ 表示边集合,如果评论者 v_i、v_j 在窗口长度为 Δt 的时间内评论相同产品的个数大于 $\kappa(\kappa\geq 1)$,则 $(L_i,L_j)\in\mathcal{E}$. $A=\{A_j\}_{j=1}^{n}$ 表示评论者能够观察到的特征集合,$A_i=(A_{i1},A_{i2},\cdots,A_{ik})$ 表示评论者 v_i 的属性集.模型的全局概率分布如下:

$$\log(Pr(l|CG))=\sum_{L_i\in L}\log(\phi_i(l_i))+\sum_{(L_i,L_j)\in E}\log(\psi_{ij}(l_i,l_j))-\log\boldsymbol{Z} \tag{47}$$

$\psi_i(l_i)$、$\psi_{ij}(l_i)$ 和 $\psi_{ij}(l_i,l_j)$ 表示 L_i 属于三类簇 $L_i\in L$、$(L_i,A_j)\in\mathcal{E}$ 和 $(L_i,L_j)\in\mathcal{E}$ 的概率. $\boldsymbol{Z}$ 是正则化因子.

$$\phi_i(l_i)=\psi_i(l_i)\prod_{(L_i,A_j)\in\mathcal{E}}\psi_{ij}(l_i) \tag{48}$$

图模型 CG 的目标是通过最大化以下公式预测评论者的类别标签,进而对所在群组进行虚假群组的判别.

$$\hat{l}=\arg\max_{l}\log(Pr(l|CG)) \tag{49}$$

基于排序思想的 GSRank 模型运用矩阵表示评论文本、评论者及产品的两两间的关系,对问题有一定的表示能力.然而方法未考虑三者间的共同作用关系,有继续探索的空间.基于图模型的 CG 方法对新加入的评论者节点有良好的扩展性,此类方法能够基于其他节点对该节点的类别做出判定.

4.3 虚假评论群组数据集

4.3.1 亚马逊虚假评论群组数据集

研究者 Mukherjee 等人[49]在亚马逊评论数据集上运用人工标注方式标注虚假评论群组.首先运用高频项集合挖掘方法(frequent itemset mining),将在不同产品评论中共现频率高的评论者集合作为候选评论群组.研究者选取了共现次数大于 3,群组成员个数大于 2 的集合作为候选评论群组,候选群组共有 7052 个.继而邀请 8 位具备领域经验的标注人员对候选群组的类别进行标注.标注人员是来自 Rediff Shopping 及 eBay.in 公司的雇员,具备领域知识及一定的识别虚假评论的能力.标注人员基于多种虚假评论特性对候选群组进行虚假性判断,如不包含告诫话语、包含大量无意义的形容词、一味地赞美以及短期内发布多条评论等.最终对其中的 2431 个评论群组产生了一致性较高的类别标注($\kappa=0.79$),其类别包括“虚假”、“真实”及“不明确”.

4.3.2 亚马逊中文评论群组数据集

研究者 Xu 等人[68]在亚马逊中国区域网站上运用高频项挖掘方法(FIM),收集了 8915 个候选评论群组.这些群组共包含 5055 个评论者.研究者标注目标是对评论者是否为合谋欺诈者进行标注,即评论者是否属于某个虚假评论群组.由于亚马逊网站会定期删除网站识别出的虚假评论,研究者间隔七个月对以上评论者所发布的评论重新检索收集,发现其中的 1822 名评论者发布了虚假评论.研究者对

剩余的评论者通过人工标注方式进行了标注. 通过以上方式对评论者的合谋欺诈身份进行了标注，数据集最终包含 1937 名合谋欺诈评论者及 3118 名非合谋欺诈者.

4.4　评价指标

目前对虚假评论群组检测的研究将该问题作为排序问题来解决. 使用的评价标准与虚假评论者检测问题的评价标准相近，具体包括：精确率(Precision)、*AUC*、*NDCG@k*(Normalized Discounted Cumulative Gain)等.

4.5　虚假评论群组检测研究小结

基于评论者的虚假评论群组检测模型从个体评论者的角度进行特征分析排序，进而通过聚类预测虚假评论群组. 该类模型自底向上由虚假评论者生成虚假评论群组，对数据集中所有评论者构成的整体网络结构没有加以考虑.

基于网络结构的虚假评论群组检测模型，自顶向下对评论者网络中的评论者节点通过相似的网络结构特征进行聚类. 该模型着重挖掘网络结构特征，对评论者的文本特征及行为特征未充分利用. 基于评论者与网络结构相结合的虚假评论群组检测模型，同时结合网络结构与评论者之间的特征相似性，相邻节点间若文本特征相似、行为特征同步一致，则属于虚假评论群组的概率较高. 该类模型综合虚假评论群组的多种影响因素，对群组特性的分析比以上两类模型更全面，能够有效提升检测性能. 我们在表 5 中对部分方法的检测性能进行了对比.

本文对虚假评论检测研究中应用的机器学习方法进行总结，如表 6 所示. 我们按照不同检测对象进行分类，从特征选择、方法复杂度、检测效果及方法

表 5　虚假评论发布者检测方法在部分公开数据集上的性能对比

序号	方法	特征	亚马逊虚假评论群组数据集	亚马逊中文评论群组数据集
1	GSRank[49]	产品集合、群组集合及群组成员集合间关系	*AUC*=95%	
2	SVM[68]	语言、个体行为、群组行为特征		*F*1=83.4%
3	KNN 与规则相结合[68]	语言、个体行为、群组行为特征		*F*1=90.9%
4	CG[68]	个体行为、群组行为特征		*F*1=91.9%

表 6　虚假评论检测现有工作所运用的机器学习方法的性能比较

方法	检测对象	特征			方法复杂度	检测效果	备注
		文本	评论者	群组			
支持向量机	文本[5,8,19,71,76]	√	√		低	高	适用于处理文本分类问题
	个体评论者[11]	√	√		低	高	
	评论群组[68]	√	√	√	低	低	对网络结构信息表示能力不足
朴素贝叶斯	文本[8,72]	√	√		低	高	模型简单、参数较少，模型学习过程对数据稀疏性不敏感；假设特征之间相互独立，在实际问题中往往不成立，对分类效果带来一定影响
逻辑回归	文本[2,10]	√			低	中	特征组合形式直观；模型对问题的拟合能力不足
稀疏相加生成模型(SAGE)	文本[7]	√			高	中	虚假评论的领域迁移能力较强
神经网络模型	文本[24,31]	√			高	高	不需要人工设计特征
语义语言模型	文本[23]	√			低	中	无监督方法，启发式规则较为简单，误判率高
“双重条件规则”模型	文本[10]	√			高	中	无监督方法，运用概率计算基于单一特征及多特征进行预测
多时间尺度检测模型	文本[52]	√	√		低	中	无监督方法，用于识别虚假评论爆发的时间窗口
双视图方法	文本[33]	√	√		高	中	半监督方法，适用于小数据集
PU 学习算法	文本[39-40,73]	√	√		高	高	半监督方法，适用于数据集包含少量正例及大量无标注数据
马尔可夫随机场	文本[43]	√	√		中	中	对评论文本的虚假性及产品的信誉度进行判断
	个体评论者[4,51-52]	√	√		中	高	对新加入的样本能够基于已有模型进行预测，扩展性较好
	评论群组[68]	√	√	√	高	高	对评论者形成的网络结构拟合能力强
类 HITS 算法	个体评论者	√	√		高	高	通过计算评论者的可靠性、评论的真实性和评价对象的信誉度三类因素之间的相关性较直观地解决该检测问题；缺点是对新加入的样本需要重新进行全网节点的计算
排序模型	评论群组[49,58]	√	√	√	中	高	方法扩展性好，阈值设定依赖人工
聚类模型	评论群组[68]	√	√	√	中	高	聚类中簇的数目较难确定

优劣总结等几个角度进行了总结.支持向量机及马尔可夫随机场在虚假评论检测的相关研究中有较为广泛的应用.支持向量机在规模较小的虚假评论有标注数据集上有较好的检测性能.马尔可夫随机场对以网状结构连接的评论者及群组有较好的数据表示能力,在对虚假评论者及虚假评论群组的检测问题中有所应用.面对有标注语料受限的问题,应用双视图模型及 PU 算法能够较好地运用少量有标注数据及大量无标注数据检测虚假评论文本.几种无监督方法基于启发式规则,对多特征的利用能力不足,有进一步改进的空间.神经网络模型能够语义表示学习辅助特征设计,在实验数据集规模较大时,有较好的检测效果.

5 研究展望

前人的研究工作为虚假评论检测任务奠定了坚实的基础,展望未来,有以下研究问题值得关注和探讨.

(1)虚假评论有标注语料规模受限的问题

现实世界中的虚假评论分布较广、数量级巨大,但难以准确标注、存在一定的误判样例.一种解决策略是运用众包方式采集.通过众包平台发布任务,召集参与者通过想象编造评论信息,构建虚假评论数据集.然而有分析表明[11]此类数据的分布特性、语法词汇特性等与现实世界的评论(如 yelp.com 网站评论)有一定的差异性,只能在一定程度上验证算法有效性.如何获得大量合理的标注数据集是该任务在研究过程中始终面临的一个难题.大量、合理的有标注数据有助于准确分析数据的特征规律,并促进虚假评论检测技术的发展.大量合理虚假评论数据集的获取或标注是一个有意义研究方向.

由于大规模有标注数据难于准确获取,如何有效利用真实世界中大量未标注评论数据是一个有意义的研究方向.目前该任务上采用的半监督方法主要有协同训练和 PU 学习,无监督学习则基于启发式方法,算法模型的探索空间很大.

在虚假评论检测研究中,面临一个领域有标注数据较丰富、而其它领域缺乏有标注数据集的问题,如何将源领域训练获得的模型应用在目标领域是一个重要的研究方向.前人工作主要集中在单一领域内部对模型算法进行训练和测试,领域迁移研究还未得到深入展开.Li 等人[7]在领域迁移问题上,尝试使用了几种浅层语义特征如 unigram、POS、LIWC 等,运用宾馆领域数据训练、在饭店和医生两个领域测试,实验发现与同领域训练分类器进行测试相比,准确率、$F1$ 等指标均下降严重,鲁棒性不强.在虚假评论检测的领域迁移问题上,需要在特征设计和机器学习方法两个层面进行更深入的探索研究.

(2)虚假评论的特征隐藏性问题

虚假评论的特征具有隐藏性,人工设计特征较为困难.研究表明[5]对虚假评论进行人工标注的评论准确率仅为 57%,这体现了人类对虚假评论识别能力较弱,也反映了通过人工进行特征设计的困难性.研究者们借助心理学及语言学的研究成果对虚假评论任务从文本、评论者的行为等角度设计、挖掘特征,在一定程度上揭示了虚假评论的语言及行为规律.特征设计需要领域专家的经验,而运用语义表示技术能够学习数据隐含的规律并将文本或结构进行向量化表示,相当于自动学习语义或关系特征.随着深度学习的近些年的蓬勃发展,基于神经网络模型的语义表示算法已在众多自然语言处理的研究任务[77-82]中验证了其算法的有效性,然而在虚假评论研究中,此类方法的运用还未得到展开.神经网络模型的通用性好于基于特征设计的检测方法,在虚假评论检测任务上的模型设计和应用值得持续的探索.

(3)虚假评论的领域多样性问题

由于评论数据涉及多个领域、由不同用户撰写,其语言、文体特点各有不同,特征具有多样性.如在酒店及饭店领域评论中,评论内容包含地点名词较多,而在医疗领域则包含的地点名词较少[7].说明同一个特征在不同领域对虚假评论的检测能力不一样.研究者们通常在同一领域对虚假评论检测模型进行训练和测试,对不同领域的通用检测模型鲜有研究.在未来对通用领域设计有效特征及检测模型是一个有意义的研究方向.

(4)虚假评论群组的动态预测问题

评论者以一定的网络拓扑结构组成群组,而群组的规模及相关特征,都伴随着群组成员的增减及行为变更而不断发生动态变化.目前的虚假评论群组的检测均基于静态网络结构.如何对群组成员不断变化、网络结构不固定的群组进行群组识别及虚假性判断,是值得深入探索的问题.在线学习技术能够对未标注的数据进行增量学习,然而在线学习技术往往无法考虑数据分布问题.将在线学习与网络结构分析相结合是对虚假评论群组进行动态预测的

一个突破口.

6　结束语

虚假评论检测是自然语言处理领域的热点问题，随着虚假评论数量的快速增长和检测技术的不断发展，该问题一直受到研究者们的关注.本文按照虚假评论文本、虚假评论者及虚假评论群组三类研究对象分别对研究方法进行分类总结，包括特征设计、模型算法的介绍和对比.进而对虚假评论检测数据集、构建方法和未来研究方向进行了总结和探讨.

致　谢　在此，向对本文在组织撰写过程中提供帮助的老师和同学们表示感谢.同时感谢编辑部和各位审稿老师的宝贵意见！

参　考　文　献

[1] López V, Río S D, Benítez J M, et al. Cost-sensitive linguistic fuzzy rule based classification systems under the MapReduce framework for imbalanced big data. Fuzzy Sets & Systems, 2015, 258: 5-38

[2] Jindal N, Liu B. Opinion spam and analysis//Proceedings of the International Conference on Web Search and Data Mining. Stanford, USA, 2008: 219-230

[3] Ott M, Cardie C, Hancock J. Estimating the prevalence of deception in online review communities. Eprint Arxiv, 2012: 201-210

[4] Fei G, Mukherjee A, Liu B, Hsu M, et al. Exploiting burstiness in reviews for review spammer detection//Proceedings of the International AAAI Conference on Weblogs and Social Media. Boston, USA, 20 13: 175-184

[5] Ott M, Choi Y, Cardie C, et al. Finding deceptive opinion spam by any stretch of the imagination//Proceedings of the Meeting of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics. Portland, USA, 2011: 309-319

[6] Crawford M, Khoshgoftaar T M, Prusa J D, et al. Survey of review spam detection using machine learning techniques. Journal of Big Data, 2015, 2(1): 1-24

[7] Li J, Ott M, Cardie C, et al. Towards a general rule for identifying deceptive opinion spam//Proceedings of the Meeting of the Association for Computational Linguistics. Baltimore, USA, 2014: 1566-1576

[8] Ott M, Cardie C, Hancock J T. Negative deceptive opinion spam//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Atlanta, USA, 2013: 497-501

[9] Guzella T S, Caminhas W M. A review of machine learning approaches to Spam filtering. Expert Systems with Applications, 2009, 36(7): 10206-10222

[10] Jindal N, Liu B, Lim E P. Finding unusual review patterns using unexpected rules//Proceedings of the ACM Conference on Information and Knowledge Management. Toronto, Canada, 2010: 1549-1552

[11] Mukherjee A, Venkataraman V, Liu B, Glance N. What yelp fake review filter might be doing//Proceedings of the International AAAI Conference on Web and Social Media. Washington, USA, 2013: 409-418

[12] Structure G &. Longman grammar of spoken and written english. Modern English Teacher, 2001, 10(4): 75-77

[13] Depaulo B M, Ansfield M E, Bell K L. Interpersonal deception theory. Communication Theory, 1996, 6(3): 297-310

[14] Rayson P, Wilson A, Leech G. Grammatical word class variation within the British National Corpus Sampler. Language & Computers, Leiden, Netherlands: Editions Rodopi BV, 2001: 295-306

[15] Shojaee S, Murad M A A, Bin Azman A, et al. Detecting deceptive reviews using lexical and syntactic features//Proceedings of the International Conference on Intelligent Systems Design and Applications. Selangor, Malaysia, 2013: 53-58

[16] Cortes C, Vapnik V. Support-vector networks. Machine Learning, 1995, 20(3): 273-297

[17] Jing Ya-Peng. Research of Deceptive Opinion Spam Recognition Based on Deep Learning [Ph. D. dissertation], East China Normal University, Shanghai, 2014(in Chinese)
(景亚鹏. 基于深度学习的欺骗性垃圾信息识别研究[博士学位论文]. 华东师范大学，上海，2014)

[18] Eisenstein J, Ahmed A, Xing E P. Sparse additive generative models of text//Proceedings of the 28th International Conference on Machine Learning. Washington, USA, 2011: 1041-1048

[19] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet allocation. Journal of Machine Learning Research, 2003, 3: 993-1022

[20] Everitt B, Rabe-Hesketh S. Generalized additive models//Everitt B S, Rabe-Hesketh S. Analyzing Medical Data Using S-PLUS. New York, USA: Springer, 2001: 590-606

[21] Li L, Ren W, Qin B, et al. Learning document representation for deceptive opinion spam detection//Proceedings of the 14th China National Conference on Chinese Computational Linguistics. Guangzhou, China, 2015: 393-404

[22] Wang Xuepeng, Liu Kang, Zhao Jun. Learning to represent review with tensor decomposition for spam detection//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Austin, USA, 2016: 866-875

[23] Lau R Y K, Liao S Y, Kwok C W, et al. Text mining and probabilistic language modeling for online review spam detection. ACM Transactions on Management Information Systems, 2011, 2(4): 1-30

[24] Abbasi A, Zhang Z, Zimbra D, et al. Detecting fake websites: The contribution of statistical learning theory. Mis Quarterly, 2010, 34(3): 435-461

[25] Pennebaker J W, Chung C K, Ireland M, et al. The development and psychometric properties of LIWC2007. Austin, USA: The University of Texas, Technical Report: 29, 2007

[26] Hancock J T, Curry L E, Woodworth S G M. On lying and being lied to: A linguistic analysis of deception in computer-mediated communication. Discourse Processes, 2008, 45(45): 1-23

[27] Mihalcea R, Strapparava C. The lie detector: Explorations in the automatic recognition of deceptive language//Proceedings of the 2009 Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing of the AFNLP. Singapore, 2009: 309-312

[28] Vrij A, Mann S, Kristen S, et al. Cues to deception and ability to detect lies as a function of police interview styles. Law & Human Behavior, 2007, 31(5): 499-518

[29] Mairesse F, Walker M A, Mehl M R, et al. Using linguistic cues for the automatic recognition of personality in conversation and text. Journal of Artificial Intelligence Research, 2007, 30(1): 457-500

[30] Zhou L, Burgoon J K, Twitchell D P, et al. A comparison of classification methods for predicting deception in computer-mediated communication. Journal of Management Information Systems, 2004, 20(4): 139-166

[31] Knapp M, Comaden M. Telling it like it isn't: A review of theory and research on deceptive communications. Human Communication Research, 1979, 5(3): 270-285

[32] Newman M L, Pennebaker J W, Berry D S, Richards J M. Lying words: Predicting deception from linguistic styles. Personality and Social Psychology Bulletin, 2003, 29(5): 665-675

[33] Li F, Huang M, Yang Y, et al. Learning to identify review spam//Proceedings of the International Joint Conference on Artificial Intelligence. Barcelona, Spain, 2011: 2488-2493

[34] El-Halees A M, Hammad A A. An approach for detecting spam in arabic opinion reviews. International Arab Journal of Information Technology, 2015, 12(1): 9-16

[35] Hai Z, Zhao P, Cheng P, et al. Deceptive review spam detection via exploiting task relatedness and unlabeled data//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Austin, USA, 2016: 1817-1826

[36] Fusilier D H, Guzman-Cabrera R, Montes-Y-Gómez M, et al. Using PU-learning to detect deceptive opinion spam//Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis. Atlanta, USA, 2013: 38-45

[37] Ren Y, Ji D, Zhang H. Positive unlabeled learning for deceptive reviews detection//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Proceeding. Doha, Qatar, 2014: 488-498

[38] Li H, Chen Z, Liu B, et al. Spotting fake reviews via collective positive-unlabeled learning//Proceedings of the IEEE International Conference on Data Mining Series. Dallas, USA, 2014: 467-475

[39] Blum A, Mitchell T. Combining labeled and unlabeled data with co-training//Proceedings of the 11th Annual Conference on Computational Learning Theory. Madison, USA, 1998: 92-100

[40] Liu B, Lee W S, Yu P S, Li X. Partially supervised classification of text documents//Proceedings of the 19th International Conference on Machine Learning. Sydney, Australia, 2002: 387-394

[41] Liu B, Dai Y, Li X, et al. Building text classifiers using positive and unlabeled examples//Proceedings of the 3rd IEEE International Conference on Data Ming. Melbourne, USA, 2003: 179-182

[42] Elkan C, Noto K. Learning classifiers from only positive and unlabeled data//Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Ming. Las Vegas, USA, 2008: 213-220

[43] Li X, Yu P S, Liu B, Ng S-K. Positive unlabeled learning for data stream classification//Proceedings of the 9th SIAM International Conference on Data Ming. Sparks, USA, 2009: 257-268

[44] Xiao Y, Liu B, Yin J, et al. Similarity-based approach for positive and unlabeled learning//Proceedings of the 22nd International Joint Conference on Artificial Intelligence. Barcelona, Spain, 2011: 1577-1582

[45] Zhang D. A simple probabilistic approach to learning from positive and unlabeled examples//Proceedings of the 5th Annual UK Workshop on Computational Intelligence. London, UK, 2005: 83-87

[46] Ren Ya-Feng, Ji Dong-Hong, Zhang Hong-Bin, et al. Deceptive reviews detection based on positive and unlabeled learning. Journal of Computer Research and Development, 2015, 52(3): 639-648(in Chinese)
(任亚峰, 姬东鸿, 张红斌等. 基于 PU 学习算法的虚假评论识别研究. 计算机研究与发展, 2015, 52(3): 639-648)

[47] Teh Y W, Jordan M I, Beal M J, Blei D M. Sharing clusters among related groups: Hierarchical Dirichlet processes//Proceedings of the Advances in Neural Information Processing Systems. Vancouver, Canada, 2004: 1385-1392

[48] Qian T, Liu B. Identifying multiple userids of the same author//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Seattle, USA, 2013: 1124-1135

[49] Mukherjee A, Liu B, Glance N. Spotting fake reviewer groups in consumer reviews//Proceedings of the International Conference on World Wide Web. Lyon, France, 2012: 191-200

[50] Ye J, Kumar S, Akoglu L. Temporal opinion spam detection by multivariate indicative signals//Proceedings of the 10th International AAAI Conference on Web and Social Media. Germany, 2016: 743-746

[51] Lim E P, Nguyen V A, Jindal N, et al. Detecting product review spammers using rating behaviors//Proceedings of the International Conference on Information & Knowledge Management. Maui, USA, 2012: 939-948

[52] Xie S, Wang G, Lin S, et al. Review spam detection via temporal pattern discovery//Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Beijing, China, 2012: 823-831

[53] Chakrabarti S. Mining the Web: Discovering Knowledge from Hypertext Data. Amsterdam, Netherlands: Elsevier, 2003

[54] Gilbert E, Karahalios K. Understanding deja reviewers// Proceedings of the ACM Conference on Computer Supported Cooperative Work. Savannah, USA, 2010: 225-228

[55] Savage D, Zhang X, Chou P, et al. Detection of opinion spam based on anomalous rating deviation. Expert Systems with Applications, 2015, 42(22): 8650-8657

[56] Danescu-Niculescu-Mizil C, Kossinets G, Kleinberg J, et al. How opinions are received by online communities: A case study on Amazon.com helpfulness votes//Proceedings of the International Conference on World Wide Web. Madrid, Spain, 2009: 141-150

[57] Mukherjee A, Kumar A, Liu B, et al. Spotting opinion spammers using behavioral footprints//Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Chicago, USA, 2013: 632-640

[58] Wu G, Greene D, Smyth B, et al. Distortion as a validation criterion in the identification of suspicious reviews//Proceedings of the Workshop on Social Media Analytics. Washington, USA, 2010: 10-13

[59] Najada H A, Zhu X. iSRD: Spam review detection with imbalanced data distributions//Proceedings of the IEEE Information Reuse and Integration. Redwood City, USA, 2014: 553-560

[60] Jindal N, Liu B. Review spam detection//Proceedings of the International Conference on World Wide Web. Banff, Canada, 2007: 1189-1190

[61] Xue H, Li F, Seo H, et al. Trust-aware review spam detection //Proceedings of the IEEE International Conference on Trust, Security and Privacy in Computing and Communications. Helsinki, Finland, 2015: 726-733

[62] Wang G, Xie S, Liu B, et al. Identify online store review spammers via social review graph. ACM Transactions on Intelligent Systems & Technology, 2012, 3(4): 1-21

[63] Wang G, Xie S, Liu B, et al. Review graph based online store review spammer detection//Proceedings of the IEEE 11th International Conference on Data Mining. Vancouver, Canada, 2011: 1242-1247

[64] Kleinberg J M. Authoritative sources in a hyperlinked environment. Journal of the ACM, 1999, 46(5): 604-632

[65] Rayana S, Akoglu L. Collective opinion spam detection: Bridging review networks and metadata//Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Sydney, Australia, 2015: 985-994

[66] Fakhraei S, Foulds J, Shashanka M, et al. Collective spammer detection in evolving multi-relational social networks// Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining. Sydney, Australia, 2015: 1769-1778

[67] Akoglu L, Chandy R, Faloutsos C. Opinion fraud detection in online reviews by network effects//Proceedings of the International AAAI Conference on Weblogs and Social Media. Cambridge, USA, 2013: 2-11

[68] Xu C, Zhang J, Chang K, et al. Uncovering collusive spammers in Chinese review websites//Proceedings of the ACM International Conference on Information and Knowledge Management. San Francisco, USA, 2013: 979-988

[69] Xu C, Zhang J. Combating product review spam campaigns via multiple heterogeneous pairwise features//Proceedings of the SIAM International Conference on Data Mining. Vancouver, Canada, 2015: 172-180

[70] Page L. The PageRank citation ranking: Bringing order to the web. Stanford Digital Libraries Working Paper, 1998, 9(1): 1-14

[71] Barabási A L, Albert R, Jeong H. Scale-free characteristics of random networks: the topology of the world-wide web. Physica A: Statistical Mechanics & Its Applications, 2000, 281(s 1-4): 69-77

[72] Benczúr A A, Csalogány K, Sarlós T, et al. SpamRank — Fully automatic link spam detection//Proceedings of the 1st International Workshop on Adversarial Information Retrieval on the Web, Co-Located with the WWW Conference. Chiba, Japan, 2005: 25-38

[73] Jiang M, Cui P, Beutel A, et al. CatchSync: Catching synchronized behavior in large directed graphs//Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining. New York, USA, 2014: 941-950

[74] Ye J, Akoglu L. Discovering opinion spammer groups by network footprints//Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Porto, Portugal, 2015: 267-282

[75] Mukherjee A, Liu B, Wang J, et al. Detecting group review spam//Proceedings of the International Conference on World Wide Web. Hyderabad, India, 2011: 93-94

[76] Taskar B, Abbeel P, Koller D. Discriminative probabilistic models for relational data//Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence. Alberta, Canada, 2002: 485-492

[77] Li J, Dan J, Hovy E. When are tree structures necessary for deep learning of representations?//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Lisbon, Portugal, 2015: 2304-2314

[78] Serban I V, Sordoni A, Bengio Y, et al. Building end-to-end dialogue systems using generative hierarchical neural network models//Proceedings of the 30th AAAI Conference on Artificial Intelligence. Germany, 2016: 3776-3783

[79] Socher R, Bauer J, Manning C D, et al. Parsing with compositional vector grammars//Proceedings of the Meeting of the Association for Computational Linguistics. Sofia, Bulgaria, 2013: 455-465

[80] Socher R, Perelygin A, Wu J Y, et al. Recursive deep models for semantic compositionality over a sentiment treebank// Proceedings of the Conference on Empirical Methods in Natural Language Processing. Seattle, USA, 2013: 1631-1642

[81] Tang D, Wei F, Yang N, et al. Learning sentiment-specific word embedding for Twitter sentiment classification// Proceedings of the Meeting of the Association for Computational Linguistics. Baltimore, USA, 2014: 1555-1565

[82] Wang P, Xu B, Xu J, et al. Semantic expansion using word embedding clustering and convolutional neural network for improving short text classification. Neurocomputing, 2016, 174(PB): 806-814

LI Lu-Yang, born in 1985, Ph. D. candidate. Her research interests include truth discovery, fake review detection and contradiction detection.

QIN Bing, born in 1968, Ph. D., professor, Ph. D. supervisor. Her research interests include text mining and text analysis.

LIU Ting, born in 1972, Ph. D., professor, Ph. D. supervisor. His research interests include social computing and natural language processing.

Background

Research on fake review detection is a fundamental study in the field of natural language processing. Many theories, models and methods of detecting fake review have been proposed and extensively studied. Although many achievements have been made in these areas, new problems are continually proposed and new challenges emerge. Especially, the arrival of big data era and the development of deep learning theory bring new opportunities and challenges for research on fake review detection. This paper clarifies the scope of fake review detection, gives a comprehensive survey on recognizing fake review, fake spammer and fake spammer group. The future research directions and new challenges are also elaborated under the current research situation.

In recent years, the authors' group has focused on the related researches. We designed a neural network based model to learn the representation of review text to detect fake review and gain good results which outperformed the state-of-art method.

This work is supported by the National High Technology Research and Development Program (863 Program) of China (No. 2015AA015407), the National Natural Science Foundation of China (No. 61632011 and No. 61370164).