计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2019,28(3):1−9 [doi: 10.15888/j.cnki.csa.006812]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

虚假评论检测技术综述①

尤苡名

(浙江理工大学 信息学院, 杭州 310018)
通讯作者: 尤苡名, E-mail: 15757126276@163.com

摘　要: 随着互联网的发展, 用户倾向于在购物、旅游、用餐之前参考线上评论. 之后, 他们也会发表评论来表达自身意见. 线上评论越来越具有价值. 评论对用户决策的重要导向作用催生了虚假评论. 虚假评论, 指用户由于利益、个人偏见等因素发布的不符合产品真实特性的评论. 这些虚假评论语言上模仿真实评论, 消费者很难识别出来. 国内外学者综合运用自然语言处理技术来研究虚假评论检测问题. 从特征工程的角度分析, 虚假评论检测方法可以分为三类: 基于语言特征和行为特征的方法、基于图结构的方法、基于表示学习的方法. 主要描述了检测的一般流程, 归纳了三类研究方法常用的特征, 比较了方法的优缺点, 并且介绍了研究常用的数据集. 最后探讨了未来研究方向.

关键词: 虚假评论; 虚假评论检测技术; 虚假评论者检测; 意见挖掘; 自然语言处理

引用格式: 尤苡名.虚假评论检测技术综述.计算机系统应用,2019,28(3):1–9. http://www.c-s-a.org.cn/1003-3254/6812.html

Survey on Review Spam Detection Techniques

YOU Yi-Ming

(School of Information Science and Technology, Zhejiang Sci-Tech University, Hangzhou 310018, China)

Abstract: With the development of the Internet, users tend to refer to online reviews before shopping, travelling, and dining. After that, they write reviews to express their own opinions. Online reviews are increasingly of great value. The significant guiding role of reviews playing in consumers' decisions has given rise to false comments, which we call review spam. The review spam refers to the comments written by users that do not meet the true characteristics of products, due to factors such as commercial profits and personal bias. Spammers imitate the writing style of true reviewers so that customers can hardly discriminate the review spam. Scholars at home and abroad use natural language processing techniques to detect review spam. From the perspective of feature engineering, review spam detection methods are divided into three types: the linguistic and behavior based, the graph based, and the representation learning based. This survey mainly describes the general process of review spam detection, summarizes feature designing of the models, and makes a comparison among three types of methods. Furthermore, the most commonly used datasets are introduced. Finally, it explores the research directions in the future.

Key words: review spam; review spam detection technique; review spammer detection; opinion mining; natural language processing

互联网的迅速发展使线上平台在人们购物、旅游、用餐、住宿等各个方面占据了重要的作用. 这些线上平台拥有大量评论. 评论信息作为桥梁联结消费者和产品, 不仅可以影响消费者的决策, 还能作为反馈来促进商家改善产品细节. 然而, 评论者个体因为利益关系、品牌偏见等原因发布了许多虚假评论. 他们甚

① 收稿时间: 2018-09-18; 修改时间: 2018-10-08; 采用时间: 2018-10-18; csa 在线出版时间: 2019-02-22

至会协同合作发布虚假评论, 组成虚假评论群组. 据调查显示[1], 美国版大众点评网站 Yelp 上欺骗性评论的比例已从 2006 年的 5% 涨至 2013 年的 20%. 虚假评论误导消费者决策, 破坏消费体验, 危害性大.

2008 年, Jindal 等[2]首次对产品虚假评论开展研究并给出虚假评论的 3 种类型:

(1) 不真实评论. 评论制造者为了提高某产品的销量, 不管产品真实的特性大肆赞美该产品, 或者为了压制某产品的销量诋毁该产品.

(2) 只关注品牌的评论. 评论者因为产品的品牌、厂商和销售商对产品带有偏见.

(3) 无关评论. 一般分为两类: 广告和其他与评论无关的文本.

由于评论内容多为短文本, 虚假评论比垃圾网页和垃圾邮件更难识别[3]. 国内外学者重点研究第一类虚假评论.

虚假评论检测难点在于找出有效的特征来更好地区分虚假评论与真实评论. 最早的时候, 研究者从评论内容提取语言特征 (例如, 词袋特征) 用于检测. 然而, 有经验的评论者编写虚假评论模仿真实评论, 所以利用评论内容识别虚假评论, 准确性不高. 于是, 研究者结合行为异常信息来提高检测准确性. 虚假评论检测另一难点在于缺少标准标注数据集评估算法性能. 研究者引入图结构, 利用评论者、评论、产品之间的关系特征, 把检测任务转为排序或者联合分类问题, 已知节点的信息通过连接的边传递到未知节点. 此类方法适用于标注数据集少的情况. da 方法检测的效率不高. 于是, 研究者利用表示学习方法让模型学习表示评论, 减少人为设计特征的繁琐性.

本文第 1 节从检测的一般流程、特征分类、检测方法三部分介绍虚假评论检测技术, 重点比较了各类方法的优缺点. 第 2 节列举了研究者们使用的合成数据集和真实世界的数据集. 第 3 节对全文进行总结, 同时探索了未来的研究方向.

1 虚假评论检测技术

1.1 检测流程

虚假评论检测的一般流程分为: 数据收集、数据预处理、特征设计、模型设计、模型评估. 数据收集指自己爬取网页数据或者下载他人整理的语料库. 数据预处理对后续的虚假评论检测性能有着很大的影响[4]. 该阶段去除了不相关信息, 并对文本进行分词、去除停用词、词性分析. 为了尽可能精确有效地表示评论, 需要对数据的特征进行分析设计, 特征设计主要包括特征提取和特征选择. 评论特征通过归一化或者规范化后输入到设计的虚假评论检测模型中. 模型评估用于检验模型的泛化性能. 常用的评估指标有: *AUC* 值、*F*1 值、准确率 *Accuracy*、精确率 *Precision*、召回率 *Recall*.

1.2 特征分类

研究中常用的特征可分为四类[5]: 评论者的语言特征、评论的语言特征、评论者的行为特征和评论的行为特征, 具体如表 1 所示. 前两类来自评论内容, 后两类由元数据产生. 这些特征是在以往的研究工作中统计出来的, 依赖于专家们对不同领域数据的经验知识.

1.3 检测方法

1.3.1 基于语言特征与行为特征的方法

基于语言学特征的方法属于早期的研究方法. 词袋特征 (unigram/bigram/trigram) 是虚假评论识别最为常用的语言特征[6-8]. Jindal 等[2]提取重复评论的 bigram 特征, 在亚马逊数据集训练回归模型, 识别只关注品牌的评论和评论文本无关的两类垃圾评论, *AUC* 值高达 90%.

Ott 等[7]仅使用 bigram 特征在合成的黄金标准数据集训练支持向量机 SVM 模型, 分类结果 *Accuracy* 达到 89.6%. Feng 等[9]利用 unigram、深层句法特征和 SVM 模型对同一合成数据集进行验证, 将 *Accuracy* 提高到 91.2%.

Li 等[10]扩充了黄金标准数据集, 研究了虚假评论检测领域迁移性问题. 研究者利用 Hotel 数据集的 Unigram 特征训练 SVM 模型和稀疏相加生成模型 (SAGE), 然后在 Restaurant 和 Doctor 数据集上测试模型. Hotel 数据和 Restaurant 数据相比有较多相似的属性, 而和 Doctor 相比相似性较少. 实验发现两个模型在 Restaurant 数据集上的分类 *Accuracy* 都能达到 75% 左右, 而在 Doctor 数据集上 *Accuracy* 只有 50% 左右. 实验说明词袋特征用于虚假评论检测领域迁移性差.

由于人工标注样例误差大, 任亚峰等[11]提出 PU 学习算法 (Positive-Unlabeled learning algorithm) 识别虚假评论. 作为半监督性学习算法, PU 算法在评论数据包含少量正例 P 和剩余全为未标注样例 U 的情况下构造分类器, 自动标记未标注样例 U. 核心是确定间谍样

例的类别标签. 该方法首先从未标注评论样例中抽取了可信负例, 利用 LDA 主题模型抽取了它们的主题分布特征, 并使用 K-Means 聚类主题分布相似的可信负例. 然后, 用 Rocchio 分类器识别出 10 个代表性正负样例, 并以代表性正负样例为基准, 混合种群性和个体性策略确定间谍样例的类别标签. 最后, 利用多核学习算法建立最终的分类器. 实验在黄金标准数据集上进行, 识别 *Accuracy* 达到 83.21%. 然而, 如果间谍样例所在子类正负样例数目相近, 并且间谍样例与代表性正负样例的相似度都不高, 算法就难以确定间谍样例的类别标签. 此外, 多核学习算法将特征映射到高维空间区分, 效率不高, 不适合处理大规模评论数据.

表 1 常用的评论、评论者的语言特征和行为特征

特征分类	特征名词	特征解释
评论者的语言特征	RL (Review Length)	平均评论长度.
	MCS (Max Content Similarity)	最大内容相似度. 比较两两评论的 cosine 相似度, 取最大值.
评论的语言特征	N-gram (unigram/bigram/trigram)	词袋特征. 对于一个文本来说, 只关注已知的词汇出现与否, 忽略其词序和词的结构. 文本中每个词的出现都是独立的. 通过计算文档中单词或词组出现的次数以及出现的频率来表示.
	POS (Part-Of-Speech)	词性分布特征. 对句子进行分词、词性标注以及统计不同词性的词出现的频率.
	Deep syntax	深层句法特征. 可以通过斯坦福句法分析器 (Stanford Parser) 分析得到.
	PP1 (ratio of 1st Person Pronouns)	评论中第一人称代词 (I, my, etc.) 占比.
	RES (Ratio of Exclamation in Sentence)	评论语句中感叹号的占比.
评论者的行为特征	MNR (Max Number of Reviews)	最大日发布评论数目.
	PR/NR (Ratio of Positive/Negtive Reviews)	评论者发布的积极/消极评论占他所有评论的比率.
	RD (Rating Deviation)	评级偏差. 评论者给的评级与其他评论者的评级是相近的. 若评论者对产品的评级与该产品的评级均值相差甚远, 则反映了该用户的异常评论行为.
	ERD (Entropy of rating distribution)	同一用户评论的评级分布熵.
	ETG (Entropy of Temporal Gaps)	同一用户连续两条评论之间的时间差Δt的熵.
	BRR (Burst Review Ratio)	突发性评论比例. 评论者短时间内发布的评论数占所有评论数的比例, 一般用于虚假评论者检测和虚假评论群组检测.
	RAVP (Ratio of Amazon Verified Purchase)	亚马逊确认购买比例. 评论的发布者购买了亚马逊的产品后, 他的评论被打上“亚马逊确认购买”的标记. 而虚假评论者往往不会购买他们评论的产品.
评论的行为特征	EXT (EXTremity of ratings)	评级的极端情况.
	ETF (Early Time Frame)	早期时间窗. 为增加产品关注度影响产品后续销售, 产品所有评论中最早发布的评论极有可能是虚假评论.
	ISR (Is Singleton Review)	如果用户仅仅发布了一条评论, 那么该评论是虚假评论的可能性非常大.

赵军等[12]提出融合情感极性和转折词的逻辑回归模型识别虚假评论. 该方法使用优势比和逐步回归变量筛选方法, 比较了 10 个文本特征和行为特征变量的显著性水平, 最后选择了 6 个对逻辑回归模型影响最为显著的特征. 实验在 Amazon 数据集上进行, 发现文本长度、情感强度和是否包含转折词的优势比最高. 将转折词和情感特征融入模型有效地提高了检测的准确性, 因为真实评论者在评论时往往比较全面. 然而, 该模型只是粗略地计算句子的情感极性, 忽略了不同副词带来的情感强度的差异. 此外, 所选择的特征中不包含时间相关的特征, 而实际上虚假评论存在爆发时间窗. 模型仍需要改进.

基于语言特征的方法应用于点评网站中的评论数据时检测效果较差. Mukherjee 等[13]使用 bigram 特征在黄金标准数据集上训练 SVM 模型, 然后将训练好的模型在 Yelp 点评网站的 Restaurant 评论数据集上测试, 仅取得 68.5% 的准确率. 研究发现[2,13,14], 将行为特征与语言学特征结合起来可以提高检测准确性. 虽然虚假评论者在语言表述上模仿真实评论者, 但是他们不能掩盖异常的评论行为.

以往的研究多次利用评论爆发性[15–18]和评论评分异常性[19–21]构建虚假评论检测模型. 评论的分布一般是随机的, 如果评论者的突发性评论集在所有评论集合中占的比例高, 那么这些评论者极有可能是虚假评

论者, 而评论者发布的突发性评论极有可能是虚假评论[22]. 然而, Li 等[23]指出, 同时出现的评论不一定都是虚假评论. 例如, 当电视广告大肆宣传产品时, 许多消费者会同时购买相同的产品, 该产品在这时间段内会产生大量的评论. 他们在大众点评的餐厅数据集上发现一种 co-bursting 行为模式, 即虚假评论者在同一小段时间内积极地对同一批餐厅发布虚假评论, 而其他时间段虚假评论者的评论行为比较消极.

Yang 等[24]发现虚假评论群组中评论者的兴趣相似 (指评论包含的方面和情感). 研究中首先找出评论内容相似的评论者集合. 然后, 利用 Author-Topic 模型[25]提取剩余评论者的评论主题分布作为评论者的兴趣向量;使用亚马逊网上商城浏览器目录接口找出同一个目录节点下并且发表时间窗为一天内的评论. 找出兴趣向量相似且评论时间窗相近的评论者作为候选者. 最终, 由三位专家判断候选者是否为虚假评论者. 实验随机选择了方法检测出的 50 名虚假评论者和 50 名真实评论者, 然后由三位专家判断真假性. 实验结果中, 虚假评论者和真实评论者的 Precision 分别为 84%、80%. 但是, 研究者并未评估所选的 3 个特征的有效性, 或者找出更多特征来提高模型分类的准确性.

将行为特征与语言特征结合可以改善虚假评论检测效果, 然而前提是需要足够的数据抽取行为信息. Wang 等[26]在 Yelp 酒店和餐厅两个领域的评论数据上研究了冷启动问题, 旨在即时检测出虚假评论者, 降低危害. 他们发现行为信息有限时, 评论长度、评论者的评级偏差、最大评论内容相似度和 bigram 特征结合较于仅使用 bigram 特征, 检测准确率提高了 5%(酒店领域), 但是 $F1$ 值降低了约 5%、召回率降低了约 19%, 而提高后的准确率也只达到 60% 左右. 这说明行为信息不够充分的情况下, 虚假评论误判率增加, 行为特征对于虚假评论的区分度有限.

1.3.2 基于图结构的方法

基于图结构的方法利用评论、评论者、产品等对象之间的关系特征, 将虚假评论者和虚假评论的检测看作联合分类或者排序问题[27]. 在该类方法中, 对象被映射为图结构中的节点, 不同对象之间的依赖关系被映射为图结构中的边. 对象与对象之间存在直接或间接的关联.

为了研究虚假评论者的检测问题, Wang 等[28]提出了异构型评论图的概念来描述评论者、评论和线上商店之间的关系. 文章采用了基于网络的算法并利用异构图各节点之间的关系来排序. 如图 1 所示, 图中存在三种类型的节点: 评论者、商店和评论. 一个评论者节点同其所写评论之间有一条边连接, 一个评论节点同该评论所关联的商店有一条边相连接. 而一个商店节点是通过评论者对该商店发表的评论与这个评论者节点间接关联.

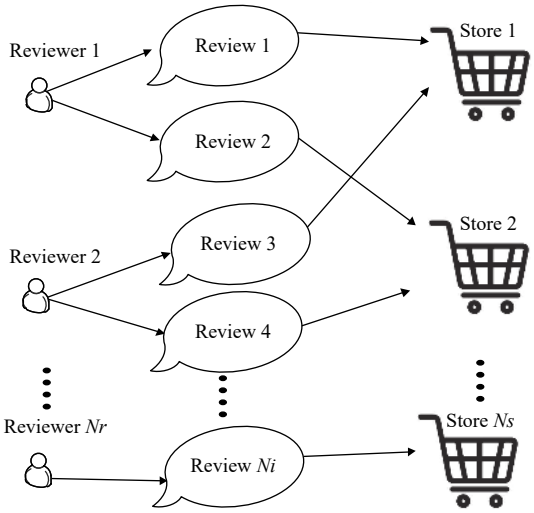

图 1 评论者-评论-商店关系图

他们还提出一个有效的迭代计算模型, 该模型采用了节点加强的方法对评论者的可信度、商店的可靠性、评论的真实性进行计算. 研究者认为评论的真实性取决于以下两点: 1) 商店的可靠性. 2) 一定时间窗内该评论与其他评论的一致性. 商店的可靠性与评论者的可信度正相关. 评论者的可信度与评论的真实性正相关. 经过多次迭代后, 各节点的信誉度将逐渐收敛, 系统也会趋于平衡. 最终, 得分较低的评论者归为虚假评论者候选人. 评论者可信度 $T(i)$ 的计算公式如公式 (1) 所示.

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

其中, H_i 表示评论者 i 的所有评论的真实性的总和. 但是, 缺点在于算法只根据计算的分数对可疑的评论者进行排序, 最终还得依靠人来评估可疑对象, 标注虚假评论者. 人工标注基于给定的规则, 在多数复杂的情况下, 还需依靠人类的直觉和大量相关信息来判断, 因此准确性有待商榷.

余传明等[29]构建个人-群体-商户模型, 量化关系特征, 迭代计算个人、群体和商户的虚假度并将其排序. 该方法构建商户-个人关系模型、商户-群体关系模型、个人-群体关系模型, 并分别计算商户和个人、商户和群体以及个人和群体的相互影响程度. 所用特征包含评论者个人行为特征、评论者群体行为特征、商家行为特征. 实验从国内大型电商平台上选取 93 家店铺、9558 个不同 IP 代表的不同评论者以及 97 804 条评论数据作为样本, 虚假评论者识别的 *Precision* 值为 92.86%, *Recall* 值为 86.47%, *F*1 值为 87.89%. 该方法不需要手动标记训练集, 消除了分类模型的训练时间, 可扩展到大型数据集. 但是, 关系模型在计算虚假度时只是简单地对特征取平均值, 忽略了不同行为特征的重要性差异.

邵珠峰等[30]构建用户之间关系的多边图模型, 计算用户的不可靠分数来识别虚假评论者. 用户节点之间存在两种类型的边. 若两个用户对同一商品评分相同或相似, 用户节点之间用支持边连接, 反之则用反对边连接. 该方法利用用户的 8 种行为特征计算用户初始特征分数, 然后归纳用户之间的支持边、反对边集合并利用 TrustRank 算法量化用户之间的关系分数, 这两部分之和为用户的不可靠分数. 最后将不可靠分数值较小的用户作为虚假评论者候选者, 邀请 3 位专业人士评估判断出虚假评论者. 缺点是, 在计算初始特征分数时, 特征权重的分配没有可靠的理论依据, 特征组合也未证明最优. 另外, 该方法凭借情感词典简单计算不同用户之间的情感特征, 分析不够全面.

Akoglu 等[31]提出 Fraudeagle 模型, 利用产品、用户、评论之间的关系识别虚假评论者. 该模型在 LBP (Loopy Belief Propagation) 算法的基础上改进. LBP 是基于信息循环传递的算法. 用户和产品映射为图节点, 评论映射为边连接节点. 对于未标记的用户, 检测过程主要分为计算分数和分组两部分. 该方法利用最大可能性概率来计算分数、标注节点. 节点的标记依赖于评论的积极或者消极情感极性. 方法的扩展性好, 运行时间与网络的大小成线性关系. 缺点在于, 加入新的节点之后就得重新迭代计算已有连接节点的概率分数. 此外, 可以考虑加入时序特征、评论文本特征来初始化节点概率分数, 提高模型识别的准确性.

Saeedreza 等[32]提出 NetSpam 模型, 利用异构型信息网络 (HIN, Heterogeneous Information Networks) 对 Yelp 和 Amazon 的评论数据集进行分类. 研究者将特征分为四类: 评论-行为特征、评论-语言特征、用户-行为特征、用户-语言特征. 该模型利用元路径量化特征重要性, 构建模型时为特征分配不同的权重. 通过实验发现, 四种类型中评论-行为特征表现最好. 选取重要的特征建模既能保证模型性能, 又降低了算法的时间复杂度. 除了基于评论者与评论的特征, 研究者指出基于产品的特征的重要性也值得分析, 但是该方法并未涉及.

1.3.3 表示学习方法

以上两类研究方法致力于设计有效的特征来区分虚假评论与真实评论, 特征设计依赖于专家的先验知识. 如果算法可以自动学习表示评论, 就可以减少人为设计特征的时间, 降低引入的噪声.

Wang 等[33]利用张量分解算法在低维向量空间表示学习评论者和产品的关系, 利用 bigram 表示评论文本, 然后将这三部分连接成一个评论整体, 作为 SVM 模型的输入. 全局特征的矢量化有效地提高了检测性能. 在 Yelp 的 Hotel 和 Restaurant 的数据集[13]上选取相同数目的虚假评论与真实评论进行实验, *F*1 值分别达到了 87.0%、89.2%, *Accuracy* 分别为 86.5%、89.9%. 但是, 该方法用 bigram 特征表示评论文本仍不够有效.

Wang 等[34]又进一步研究了虚假评论是语言异常还是行为异常的问题. 针对虚假评论的现状, 即有些评论者富有经验, 在发表评论时善于伪装, 此时主要利用虚假评论者异常的行为区分虚假评论;另一些评论者则相反, 评论中往往包含更多的语气词、情感词, 体现出较强的情感强度, 所以只要利用语言特征就容易区分出虚假评论. 研究方法利用 MLP 多层感知机学习行为特征向量, 利用 CNN 卷积神经网络学习语言特征向量, 并引入 Attention 机制动态学习行为特征和语言特征的权重. 最终相比于 Mukherjee 等[13]使用现成的 SVM 分类模型, *F*1 值提高了 1.5%, *Accuracy* 提高了 1.2%. 这说明了现有模型对虚假评论检测效果仍然有限. 另外, 相比于研究者此前工作[33], *F*1 值、*Accuracy* 分别提高了 1.9% 和 2.3%. Attention 机制有效地区分了虚假评论属于语言异常或是行为异常. 至今为止, 该方法在 Yelp 评论数据集上检测的 *F*1 值和 *Accuracy* 值最优. 然而遗憾的是, 研究者未在其它实验数据上验证所提算法的健壮性.

张李义等[35]结合深度置信网络 DBN 和模糊集识

别淘宝的虚假交易. 该方法利用用户的历史评论和交易记录提取表示用户行为的 12 个特征. 首先, 无监督地训练每一层受限玻尔兹曼机网络. 然后, 根据输入特征向量和顶层降维后传递的重构特征向量之间的误差对整个 DBN 网络进行有监督反馈微调. 接着, 采用模糊集描述用户“是刷客”或者“不是刷客”的隶属度. 最后, 将识别出的“刷客”的交易认定为虚假交易. 实验结果中准确率、精确率、召回率、*F*1 值分别达到 89%, 84.21%, 96% 和 89.72%. DBN 作为深层网络学习结构, 能够学习抽象特征, 弱化浅层结构的错误特征, 从而缓解过拟合现象, 提高模型分类效果. 局限性在于, 该方法分别选取了 100 名“刷客”和正常用户进行算法验证, 相比于电商平台海量的用户, 数据量过少.

Dong 等[36]提出端到端 (end-to-end) 混合神经网络和随机森林的模型来识别虚假评论. 随机森林作为集成学习算法, 在训练时能防止每一决策树过拟合. 该方法利用 Autoencoder 算法自动表示评论特征, 作为随机森林的输入. 该方法巧妙地结合了深度学习和传统分类模型, 为虚假评论检测提供了新思路. 在 Amazon 数据集[37]上实验, *Accuracy* 达到 96%. 但是, 该方法需要设置合适的参数平衡时间消耗和预测性能的关系. 这需要反复实验调整. 此外, Autoencoder 算法也被用于微博垃圾评论检测[38].

1.3.4 小结

基于语言学特征和行为特征的方法使用的模型一般较为简单, 检测的效果相对较好, 但是特征设计过程耗时且具有挑战性. 不同数据集的数据稀疏程度、涉及的领域、语言的表述、评论者的关注面不同. 所以, 针对不同的数据集, 需选取不同的特征进行实验. 另外, 特征设计一般依靠专家的经验, 而专家们的经验也不完全可靠.

基于图结构的方法利用了评论、评论者、产品和商店之间的网络关系, 使用传播算法、迭代算法等计算节点的分数. 这类方法适用于标注数据稀少或者无标注数据的情况. 在虚假评论检测问题上, 优点是可以不依赖于人工标注数据, 扩展性好. 缺点是计算信誉度时利用的规则往往比较单一, 新加入的节点影响已有节点的分数, 所以需要重新迭代计算已有节点的分数. 该类方法适用的网络规模不宜过大, 而且检测效果还有待提升.

以上两类方法用到的特征通过统计得到, 而表示学习方法能自动学习表示评论, 既能提高实验效率又能提升检测效果. 虚假评论者为了躲避网站算法检测, 可能会增加评论的细节信息, 或者利用账号积攒信用后发布虚假评论. 可见虚假评论的语言特征与行为特征是动态变化的, 不可预知的. 表示学习方法不需要依赖经验设计特征, 因此鲁棒性好. 这类方法作为最新的研究趋势, 检测效果优于传统的方法, 然而这方面的研究较少而且不够深入.

三类方法的比较具体见表 2.

表 2 三类方法的特点比较

方法	经典模型	适用范围	复杂度	优点	缺点
基于语言特征与行为特征的方法	支持向量机、逻辑回归、朴素贝叶斯等	标注数据, 数据规模小, 分类问题, 监督性/半监督性	低	利用现有的分类器模型, 模型简单.	构建特征工程耗时繁琐, 领域迁移性差
基于图结构的方法	HIN 异构信息网络、LBP 信息循环传递、隐马尔科夫等	无标注数据/标注数据少, 数据规模中等. 排序问题, 无监督/半监督	较高	拟合真实世界数据分布关系, 利用网络连接关系识别出伪装的欺骗性评论, 扩展性好.	规则单一, 反复迭代
基于表示学习的方法	CNN 卷积神经网络、深度置信网络 DBN、Autoencoder 自编码器、张量分解等	标注数据少, 整体数据规模大, 半监督性	高	自动学习表示评论, 减少构建特征工程的时间, 鲁棒性好.	参数设定困难, 模型容易过拟合, 解释性差

2 数据集

研究者们不但致力于选择有效的特征表示评论/评论者, 寻找合适的模型提高检测效果, 而且探索研究多领域数据. 但是, 虚假评论检测研究主要问题是: 缺少标准标注数据集来评估算法性能. 目前, 研究者们主要利用众包平台构造的评论数据或者真实世界点评网站的评论数据.

2.1 众包平台构造的数据集

众包平台通过向员工分配需求任务, 依靠人类的智慧来完成计算机还不能完成的任务. 例如, 从许多照

片中挑出最棒的商店前台的照片, 编写产品描述性评论, 或者区分出音乐 CD 封面上的歌手等[39].

Ott 等[7]利用亚马逊众包平台获取黄金标准数据集, 这是唯一公开可用的数据集. 研究者通过向线上人员支付 1$酬金令他们对 20 个受欢迎的芝加哥酒店构建想象型的积极评论, 共收集了 400 条虚假评论. 此外, 研究者在 TripAdvisor.com 上收集了这 20 家酒店的 400 条积极评论作为真实评论. 之后, Li 等[10]为了研究分类器在不同领域的迁移性能, 扩充了这 800 条评论数据集, 构造了跨酒店、餐厅、医院 3 个领域的黄金标准数据集. 该黄金标准数据集包含了 3 种类型的评论: 领域专家的虚假评论, 众包平台的虚假评论以及消费者的真实评论. 实验结果表明, 酒店评论数据集训练成的分类模型在餐厅和医院评论数据上分类效果不佳.

众包平台的员工并未刻意模仿真实评论的表述, 构造出的虚假评论和现实世界中的评论存在着较大差异.

2.2 点评网站的数据集

点评网站一般有自己的虚假评论过滤算法, 这些过滤算法是商业机密, 不向外部开放. 表 3 概括了来源于点评网站的研究常用数据集. 其中, Yelp 评论数据集[13]作为近似标准标注数据集被广泛用于虚假评论检测的学术研究中. 而 Amazon 评论数据集[37]由于数据量大极具研究价值, 主要应用于情感分析、观点挖掘、产品推荐、虚假评论检测等各个领域.

表 3 点评网站评论数据集

数据集	语言	包含领域	获取途径	数据量	采集时间	数据来源	标注方式
Dianping[23]	中文	上海 500 家餐厅	未公开	3523 条虚假评论, 6242 条未过滤评论	2011.11.1 –2013.11.28	Dianping.com	网站过滤算法
Yelp[5,32] YelpChi	英文	芝加哥酒店餐厅	未公开	酒店: 802 条虚假评论, 4872 条未过滤评论. 餐厅: 8368 条虚假评论, 50 149 条未过滤评论.	2013	Yelp.com	网站过滤算法
YelpNYC		纽约餐厅	未公开	923 家餐厅 359 052 条评论	2015		
YelpZip		纽约餐厅	未公开	5044 家餐厅 608 598 条评论	2015		
TripAdivisor[40]	英文	宾馆	http://mlg.ucd.ie/datasets/trip	3 万条评论	2010	TripAdivisor.com	未标注
Amazon[2]	英文	书籍、音乐、DVD/VHS 等产品	http://liu.cs.uic.edu/download/data/	5838 032 条评论, 1195 133 个产品	2006.6	Amazon.com	未标注
Resellerrating[28]	英文	线上商店	未公开	408470 条评论	2010.10.6	resellerratings.com	未标注
SWM[31]	英文	娱乐类别下的评论 (如游戏、电影、新闻等)	未公开	1132 373 条评论	2012.6	匿名的线上商店数据库	未标注

3 总结展望

近年来, 线上消费者在做出决策前都会参考商业网站的产品评论. 真实可靠的评论既能改善消费者体验, 也能促进商家良性竞争. 本文主要概括了研究常用的四类特征, 总结了国内外研究者提出的虚假评论检测方法, 并从特征工程的角度对比了基于语言特征和行为特征的方法、基于图结构的方法、基于表示学习方法的优缺点, 最后列举了研究中使用的数据集. 从现阶段的检测技术来看, 虚假评论检测仍有很大的探索空间, 具体归纳为以下 4 点:

(1) 针对来自不同领域的数据集, 研究者们需要选取不同的特征来构建分类器, 重复特征选择这一工作. 这说明未来需要探索跨领域实验来优化特征选择的过程, 减少重复性的人工操作. 此外, 最优的特征选择也是未来的探索方向.

(2) 真实世界中虚假评论数据与真实评论数据不平衡, 不平衡的数据训练出的模型效果较差. 以往的研究通常利用采样达到数据平衡. 然而, 训练的模型在测试自然分布的数据集时检测效果下降. 未来可以探索更多适用于真实世界中不平衡数据的技术.

(3) 公开的真实评论网站的数据集较少, 以往的研究大多使用了人工构造的数据集. 但研究证实, 经人工构造的数据集训练出的分类器在对真实世界的评论数据进行分类时效果不理想[13]. 所以, 可以进一步探索如

何利用真实世界大量未标注数据来获取合理的虚假评论数据集.

(4) 虚假评论的冷启动问题. Wang 等[26]针对这个未被前人探索过的问题, 提出了一个基于图结构与 CNN 卷积神经网络的模型. 评论的真实性越早判别, 造成的不利影响越小. 新用户只发布一条虚假评论时, 如何利用先验知识准确地判别评论的真实性具有重大意义. 未来可以探索更多有效的检测模型.

参考文献

1 Luca M, Zervas G. Fake it till you make it: Reputation, competition, and Yelp review fraud. Management Science, 2016, 62(12): 3412–3427. [doi: 10.1287/mnsc.2015.2304]

2 Jindal N, Liu B. Opinion spam and analysis. Proceedings of 2008 International Conference on Web Search and Data Mining. Palo Alto, CA, USA. 2008. 219–230.

3 Khurshid F, Zhu Y, Yohannese CW, *et al*. Recital of supervised learning on review spam detection: An empirical analysis. Proceedings of the 2017 12th International Conference on Intelligent Systems and Knowledge Engineering. Nanjing, China. 2017. 1–6.

4 Etaiwi W, Naymat G. The impact of applying different preprocessing steps on review spam detection. Procedia Computer Science, 2017, 113: 273–279. [doi: 10.1016/j.procs.2017.08.368]

5 Rayana S, Akoglu L. Collective opinion spam detection: Bridging review networks and metadata. Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Sydney, NSW, Australia. 2015. 985–994.

6 Ott M, Cardie C, Hancock J. Estimating the prevalence of deception in online review communities. Proceedings of the 21st International Conference on World Wide Web. Lyon, France. 2012. 201–210.

7 Ott M, Choi Y, Cardie C, *et al*. Finding deceptive opinion spam by any stretch of the imagination. Meeting of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2011. 309–319.

8 Xu C, Zhang J. Combating product review spam campaigns via multiple heterogeneous pairwise features. Proceedings of the 2015 SIAM International Conference on Data Mining. Vancouver, BC, Canada. 2015. 172–180.

9 Feng S, Banerjee R, Choi Y. Syntactic stylometry for deception detection. Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics. Jeju, the Republic of Korea. 2012. 171–175.

10 Li JW, Ott M, Cardie C, *et al*. Towards a general rule for identifying deceptive opinion spam. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. Baltimore, MD, USA. 2014. 1566–1576.

11 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.

12 赵军, 王红. 融合情感极性和逻辑回归的虚假评论检测方法. 智能系统学报, 2016, 11(3): 336–342.

13 Mukherjee A, Venkataraman V, Liu B, *et al*. What Yelp fake review filter might be doing? Proceedings of the 7th International Conference on Weblogs and Social Media. Palo Alto, CA, USA. 2013. 409–418.

14 Mukherjee A, Venkataraman V, Liu B, *et al*. Fake review detection: Classification and analysis of real and pseudo reviews. Technical Report UIC-CS-2013-03. Chicago: University of Illinois at Chicago, 2013.

15 Mukherjee A, Liu B, Glance N. Spotting fake reviewer groups in consumer reviews. Proceedings of the 21st International Conference on World Wide Web. Lyon, France. 2012. 191–200.

16 Ye J, Akoglu L. Discovering opinion spammer groups by network footprints. Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, Cham. 2015. 267-282.

17 Xu C, Zhang J, Chang K, *et al*. Uncovering collusive spammers in Chinese review websites. Proceedings of the 22nd ACM International Conference on Information &Knowledge Management. ACM. 2013. 979–988.

18 Xu C, Zhang J. Towards collusive fraud detection in online reviews. Proceedings of 2015 IEEE International Conference on Data Mining. Atlantic City, NJ, USA. 2015. 1051–1056.

19 Viswanath B, Bashir MA, Crovella M, *et al*. Towards detecting anomalous user behavior in online social networks. Proceedings of the 23rd USENIX Security Symposium. San Diego, CA, USA. 2014. 223–238.

20 Lim E P, Nguyen V A, Jindal N, *et al*. Detecting product review spammers using rating behaviors. Proceedings of the 19th ACM International Conference on Information and Knowledge Management. ACM. 2010. 939–948.

21 Xie SH, Wang G, Lin SY, *et al*. Review spam detection via temporal pattern discovery. Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Beijing, China. 2012. 823–831.

22 Fei GL, Mukherjee A, Liu B, *et al*. Exploiting burstiness in

reviews for review spammer detection. Proceedings of the 7th International AAAI Conference on Weblogs and Social Media. Ann Arbor, MI, USA. 2013. 175–184.

23 Li HY, Fei GL, Wang S, *et al*. Bimodal distribution and co-bursting in review spam detection. Proceedings of the 26th International Conference on World Wide Web. Perth, Australia. 2017. 1063–1072.

24 Yang M, Lu ZY, Chen XJ, *et al*. Detecting review spammer groups. Proceedings of the 31th AAAI Conference on Artificial Intelligence. San Francisco, CA,USA. 2017. 5011–5012.

25 Rosen-Zvi M, Griffiths T, Steyvers M, *et al*. The author-topic model for authors and documents. Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence. Banff, Alberta, Canada. 2004. 487–494.

26 Wang XP, Liu K, Zhao J. Handling cold-start problem in review spam detection by jointly embedding texts and behaviors. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver, BC, Canada. 2017. 366–376.

27 Dewang RK, Singh AK. State-of-art approaches for review spammer detection: A survey. Journal of Intelligent Information Systems, 2018, 50(2): 231–264. [doi: 10.1007/s10844-017-0454-7]

28 Wang G, Xie SH, Liu B, *et al*. Review graph based online store review spammer detection. Proceedings of the 2011 IEEE 11th International Conference on Data Mining. Vancouver, BC, Canada. 2011. 1242–1247.

29 余传明, 冯博琳, 左宇恒, 等. 基于个人-群体-商户关系模型的虚假评论识别研究. 北京大学学报 (自然科学版), 2017, 53(2): 262–272.

30 邵珠峰, 姬东鸿. 基于情感特征和用户关系的虚假评论者的识别. 计算机应用与软件, 2016, 33(5): 158–161, 172. [doi: 10.3969/j.issn.1000-386x.2016.05.039]

31 Akoglu L, Chandy R, Faloutsos C. Opinion fraud detection in online reviews by network effects. Proceedings of the 7th international AAAI Conference on Weblogs and Social Media. Ann Arbor, MI, USA. 2013. 2–11.

32 Shehnepoor S, Salehi M, Farahbakhsh R, *et al*. NetSpam: A network-based spam detection framework for reviews in online social media. IEEE Transactions on Information Forensics and Security, 2017, 12(7): 1585–1595. [doi: 10.1109/TIFS.2017.2675361]

33 Wang XP, Liu K, He SZ, *et al*. Learning to represent review with tensor decomposition for spam detection. Proceedings of 2016 Conference on Empirical Methods in Natural Language Processing. Austin, TX, USA. 2016. 866–875.

34 Wang XP, Liu K, Zhao J. Detecting deceptive review spam via attention-based neural networks. In: Huang XJ, Jiang J, Zhao DY, *et al*., eds. Natural Language Processing and Chinese Computing. Cham: Springer, 2018. 866–876.

35 张李义, 刘畅. 结合深度置信网络和模糊集的虚假交易识别研究. 现代图书情报技术, 2016, 32(1): 32–39.

36 Dong MQ, Yao LN, Wang XZ, *et al*. Opinion fraud detection via neural autoencoder decision forest. arXiv: 1805.03379, 2018.

37 He RN, McAuley J. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. Proceedings of the 25th International Conference on World Wide Web. Montréal, Québec, Canada. 2016. 507–517.

38 Wang BH, Huang JL, Zheng HH, *et al*. Semi-supervised recursive autoencoders for social review spam detection. Proceedings of the 2016 12th International Conference on Computational Intelligence and Security. Wuxi, China. 2016. 116–119.

39 Amazon mechanical turk. https://en.wikipedia.org/wiki/Amazon_Mechanical_Turk. (2018-07-19)[2018-08-02]

40 Wu GY, Greene D, Smyth B, *et al*. Distortion as a validation criterion in the identification of suspicious reviews. Proceedings of the 1st Workshop on Social Media Analytics. Washington, DC, USA. 2010. 10–13.