文章编号：2095-2163(2023)10-0052-04 中图分类号：TP391 文献标志码：A

虚假评论特征提取检测技术研究

张锏予
（沈阳理工大学 信息科学与工程学院，沈阳 110158）

摘 要：随着互联网技术的发展和网上购物的常态化，当前存在诸多的网上购物虚假评论问题，本文对虚假评论特征提取技术展开了研究。首先，对国内外的评论特征提取及检测技术进行了归纳，将评论特征提取的方法分为 3 种方式，分别为基于传统方法、深度学习方法和机器学习方法；其次，针对 Yelp 店铺数据集特征提取，利用多种机器学习分类器比较融合方法，分析了不同分类器对此数据集虚假评论特征检测的优劣。

关键词：虚假评论；特征提取；机器学习分类器

Research on feature extraction and detection of fake comment

ZHANG Jianyu
（School of Information Science and Engineering, Shenyang Ligong University, Shenyang 110158, China）

【Abstract】 With the gradual development of Internet technology and online shopping, aiming at the problem of fake comments in current online shopping, this paper studies the feature extraction technology of fake comments. Firstly, the detection technology of comment feature extraction at home and abroad is summarized. Then a series feature extraction experiments are conducted on Yelp store dataset. We test a variety of machine learning classifiers and compare the fusion of different methods. The advantages and disadvantages of different classifiers in detecting fake comments are analyzed.

【Key words】 fake comment; feature extraction; machine learning classifier

0 引 言

随着电子商务与互联网技术的迅猛发展，消费者的消费方式也从传统的线下消费转移到了线上购物。而消费者为选择合适的商品，会参考商品的用户评论信息。消费者判断相关商家的诚信度和商品质量的好坏会受到虚假评论的影响，这些虚假的评论信息会诱导消费者对一些不符合实际的商家服务、商品价值、商品质量等进行选择，严重干扰了消费者的购物选择，扰乱了网络电商的运营。

针对网上购物场景中的虚假评论，本文采用评论特征提取检测技术，确定虚假评论中的标识文本内容，将虚假评论与其他真实评论区分开。随着机器学习的应用与发展，虚假评论特征提取检测技术的发展与日俱进[1]。但由于虚假评论是由商家或企业利用大量水军发布的，而水军可以通过多个账号进行评价，留下的痕迹难以捕捉，目前没有先进的技术可用于检测这些虚假评论，所以高精确率、低成本要求、方便客户操作和有效筛选的虚假评论特征提取技术的研究是未来的重点研究方向。

1 相关工作

虚假信息泛滥，品牌诚信对建立消费者信任至关重要，置信度有可能直接转化为利润。检测过滤出虚假评论，对于确保在线评论反馈系统的完整性、可靠性至关重要。目前主要有 2 种解决方法：一种是基于传统方法的特征提取检测；另一种是基于深度学习的特征提取检测方法。

1.1 基于传统方法的特征提取

基于传统的提取评论方法是根据事实情况，手动的核对虚假信息中的虚假内容及观点，通过将信息表达与核实的真实表达比较，判断评论信息的准确度。而手动核对虚假信息又可分为两种方式，一种是基于专家的手动核查，通过对评论的整段评价，对详述内容的可靠性评级，对词句、语法的正确表达进行筛选、评价，保证评论提取的准确率，但是当评论检测数量激增时，准确性会大打折扣；另一种是众包的方法，利用群众的数量优势对评论进行提取筛

作者简介：张锏予（1998-），女，硕士研究生，主要研究方向：大数据与智能信息处理技术研究。

收稿日期：2022-10-04

查，可以获得较低的成本付出，但是人工方法检测虚假评论的精度仅为 57%，评论提取的准确率不高是尚未解决的问题[2]。

1.2　基于深度学习的特征提取

随着深度学习算法的不断发展，深度学习算法也应用在特征提取领域[3]。卷积神经网络（CNN）被应用在矩阵分解模型中，通过从评论中提取需要的特征量，对评论进行评分预测，并通过概率矩阵分解达到特征提取的效果，但模型无法验证评论特征的重要程度。Trans-Nets[4] 通过拓展，构建了基于并行神经网络的 Deep-Conn 双塔结构模型，将隐藏层的引入作为评论描述和商品实际特点的转化；而 D-ATTN（Dual Attention model）模型以及 NARRE（Neural Attentional Regression model with Review-level Explanations）模型在 Deep-Conn 模型的基础上引入注意力机制，可以轻松的抓到评论文本中的中的关键要素及信息[5-6]；DAML 模型集成了交互注意力机制，在捕获用户和商品特征后，展现用户和特征评论的关联，特征交互由神经因子分解机完成[7]。

1.3　基于机器学习的特征提取

基于机器学习提取特征包含 4 个部分，分别是：基于文本内容重复评论提取特征；基于评论人属性与行为提取特征；基于评论主观性的特征提取；基于特征融合的方法。

1.3.1　基于文本内容重复评论提取特征

对于大部分发布虚假评论的用户而言，不论评论的是同类型商品还是不同类型商品，虚假评价内容都具有极高相似度[8]。当某些评论里的内容和语言表达出现一定程度的相似或覆盖时，就可将相似的部分作为特征提取的训练集，对训练集进行虚假评论特征提取训练。

1.3.2　基于评论人属性与行为提取特征

Hussain 开发了一个评论图来捕捉评论、评论者和商店之间的互动，评论的真实性是可以计算的，但这种方法没有使用任何评论文本信息[9]。相比之下，Wang[10] 提出的方法仅基于文本特征，研究了几个特征类别对垃圾评论识别的影响，包括打分时间、内容、情感、产品或个人资料特征。

1.3.3　基于评论主观性的特征提取

从评论主观性角度分析，需要引入情感特征。如果评论中的表达显得过于吹捧或者诋毁，则很可能是虚假的无意义评论，因此可以通过情感分析体现评论内容的主观性和褒贬性。在现有研究中，一般利用情感词汇的极性对文本的情感倾向进行评价，目前主要有利用情感词数或利用情感词典计算情感强度的加权得分两种度量方法。

1.3.4　基于特征融合的方法

在检测虚假评论时，不仅需要提取关于评论内容的特征，还需要提取其他特征，如评论者信息、评论者关注数量、收藏商品等来辅助检测。

2　多机器学习分类器比较

由于虚假评论与真实评论特征散乱，欺诈隐蔽性较强，无明显分布区分度，故而需要借助多种机器学习算法，进行有监督检测学习。当前使用较多的机器学习分类器包括 K 邻近（KNN）、支持向量机（SVM）、朴素贝叶斯（NB）、决策树（DT）等等。

2.1　K 邻近（KNN）

K 邻近算法分类是测量文本特征中不同特征值互相的距离。假设特征空间中样本的 K 个最邻近的都同属一种类型，那么在特征空间中的这个样本也属于这个类型。KNN 算法具有很多优点，操作简单、理论清晰且无需参数支持等。在多种分类要求的问题上，KNN 可提供更高的效率及准确度，但是 KNN 算法对样本数量的要求较高，需要使用很大的算力，内存消耗大。

2.2　支持向量机（SVM）

支持向量机通过给定系统的训练样本集，使得系统在训练样本集中找到无数个超平面，区分不同类型的样本。通过超平面做分类的支持向量机无需将样本集中的所有样本进行计算，可以提高运算效率，节省内存。支持向量机的缺点是在计算时需要将一些没有规章且维度较低的数据，在核函数的映射下，映射到高维空间，且使用超平面将样本区分，较为复杂。

2.3　朴素贝叶斯（NB）

朴素贝叶斯算法是贝叶斯公式和条件独立假设方法的结合应用。当文本中的某些特征项不能通过直接统计获得，则可以使用概率公式进行转换，通过加强的假设，将概率进行乘法运算，从而得到对应的属性概率。

朴素贝叶斯算法可以设置先验概率，通过一系列简单的数学计算就可以实现，大大节省了内存和运算时间，缺点是仅适用于文本样本，且样本特征相互独立。

2.4　决策树（DT）

决策树是一种基本的机器学习模型，可以用树

形图表示的树结构,以此表示各个属性与其对象值之间的映射关系。在决策树的整体结构中,每个叶节点代表一个待预测的标签类型,每个内部节点对应于一个属性,如果某些节点具有与之相对应的属性,则二者之间可能存在分支。针对提取的特征应用决策树进行预测,通过递归分割过程,直至实现所有的子集包含一样的目标量,但决策树算法在训练过程中时间成本较高。

2.5 融合分类器(LGB)

轻量级梯度提升分类器 LGB 在不损害准确率的条件下加快 GBDT 模型的训练速度,且占用内存更少,主要目的是利用弱分类器(决策树)迭代训练以得到最优模型,广泛应用于分类、预测等领域。

3 实验验证与结果分析

3.1 数据集

本文使用公开可用的 Yelp 数据集,该数据集应用广泛且声誉良好,采用 Yelpzip 子集进行实验。该数据集中 86.78%的数据被标记为真实评论,13.22%为虚假评论,显然非常不平衡。因此,在建立相应的分类模型之前,采用下采样算法平衡数据集,减少分类器的识别误差。这种方法优点是减少数据中的噪声点,避免过拟合,缺点是减少了可学习的数据量。

3.2 实验特征提取

Salminen J[11] 分析得出在虚假评论检测任务中,行为特征比单一文本特征更加有效。故本文选用基于特征融合的方法提取 Yelp 酒店和餐厅领域中行为和文本特征,并分析其有效性。

(1)活跃时间窗(AW):虚假评论者很可能在短时间内进行评论,通常不是长期活跃的成员。将该评论者的最后一次和第一次评论的时间戳之差作为活动窗口,检测每一位评论者在指定时间窗内的活跃度。大多数的虚假评论者的活跃时间为 2 个月,而真实评论者的活跃时间少于 10 个月。

(2)最大评论数(MNR):表示一天内的最大评论数。在数据中,约三分之一的虚假评论人在一天内发布了所有的评论,大部分的虚假评论人每天写 6 条或更多的评论,而真实评论者的日评论率非常适中。

(3)评论计数(RC):表示评论者的评论数量。大多数的虚假评论者发布评论数量在 11 条之内,而半数的真实评论者评论数量超过 40 条。虚假评论者和真实评论者评论数量有明显的区分。

(4)正面评价百分比(PR):正面评价(高于 3 分)占全部评价的百分比越高越可疑。大多数的虚假评论者的目标是提升企业口碑,正面评级较多。而在现实生活中,由于评价标准不同,真实评论者的评级表现出均衡的分布趋势,不同范围的评论者拥有不同比例的正面评论。

(5)评论长度(RL):大多数虚假评论的平均评论长度限制在 135 个单词以内,而大多数真实用户的平均评论字符长度高于 200 个字符。

(6)评论人偏差(RD):虚假评论者偏离一般消费者评级共识的数量。为了测量评论者的偏差,首先计算一个评论人与同一产品的其他评论人之间的绝对评分偏差;其次,取其所有评论的所有评级偏差的平均值,计算该评论者的平均偏差。在满分为 5 的尺度上,偏差可以从 0~4。大多数真实评论人在五星尺度上的绝对偏差为 0.6,这表明真实评论人与其他真实评论人对产品有评级共识,而大多数虚假评论者与真实评论者的评级偏差较大。

(7)最大内容相似度(MCS):即同一评论者的任意两条评论内容的余弦相似度。大多数真实评论人在评论中几乎没有相似度(以 0.16 余弦相似度为界);而大多数的虚假评论者在评论中有较高相似度。

通过融合上述 7 种互不相关的有效特征,可提高虚假评论检测水平。信息融合越全面,特征提取效率越高。

3.3 实验结果分析

由于消费者在消费前习惯于参考平台的最新消费评价信息,使得虚假评论往往在某一时间窗内呈爆发趋势。选取 Yelpzip 子集近两年的评论数据,并随机选取其中 80%数据集作为训练集,其余作为测试集,采用交叉验证法,比较不同分类模型的预测性能优劣,分类结果见表 1。从召回率来看,LGB 模型是检测效果最佳的模型。

表 1 交叉验证机器学习模型分类结果

Tab. 1 Classification results of different machine learning models

%

模型	*AUC*	精度	召回率	*F*1 值
KNN	52.87	85.88	95.69	90.52
SVM	52.74	86.04	97.37	91.35
NB	52.17	84.26	94.88	89.25
DT	52.69	86.29	95.19	90.52
LGB	50.22	85.06	99.79	91.84

AUC(Area Under the Curve of ROC)是评估分类器性能的主流数值指标,能够很好地平衡使用不

同概率阈值的预测模型的真阳性率和假阳性率，所以针对严重不均衡的评论数据集，往往将高 *AUC* 值作为预测性能的首要评价指标。将下采样法应用于 Yelp 数据集，机器学习模型分类结果见表 2。各个分类器模型的 *AUC* 值均有所提高，LGB 模型增长最为显著，证实了基于分类器融合的有监督方法在虚假评论检测中具有较好效果，但需要在召回率和精度之间做出权衡。此外，单纯基于文本重复、评论人行为和评论主观属性中一方面进行特征提取的检测效果远低于多特征融合特征提取。因此，虚假评论检测精度与互不重叠的有效文本特征数呈正相关。

表 2　下采样后机器学习模型分类结果

Tab. 2　Classification results after downsampling　%

模型	*AUC*	精度	召回率	*F*1 值
KNN	55.89	87.84	61.54	72.37
SVM	58.17	88.92	60.30	71.86
NB	53.69	92.51	15.77	26.94
DT	58.41	88.58	58.54	70.49
LGB	57.69	88.43	60.01	71.49

4　结束语

本文针对 Yelp 数据集中的已标注虚假评论，提取虚假评论的文本特征和行为特征，运用多种机器学习比较融合的方法，对虚假评论进行有监督机器学习分类。实验结果表明，Yelpzip 数据集极不均衡且虚假评论特征隐蔽性强，有监督方法在虚假评论检测中具有一定效果；提出利用下采样法在分类检测过程中平衡检测精度和召回率；有监督方法在实际应用中取得了较好效果，也可为下一步设计基于在线虚假评论特征自动提取检测技术方法提供参考。

参考文献

[1] MOHAWESH R, XU S, TRAN S N, et al. Fake reviews detection: A survey[J]. IEEE Access, 2021, 9: 65771-65802.

[2] PLOTKINA D, MUNZEL A, PALLUD J. Illusions of truth—Experimental insights into human and algorithmic detections of fake online reviews[J]. Journal of Business Research, 2020, 109: 511-523.

[3] ZHANG S, YAO L, SUN A, et al. Deeplearning based recommender system: A survey and new perspectives[J]. ACM computing surveys (CSUR), 2019, 52(1): 1-38.

[4] ZHENG L, NOROOZI V, YU P S. Joint deep modeling of users and items using reviews for recommendation[C]//Proceedings of the tenth ACM international conference on web search and data mining. 2017: 425-434.

[5] SEO S, HUANG J, YANG H, et al. Interpretable convolutional neural networks with dual local and global attention for review rating prediction[C]//Proceedings of the eleventh ACM conference on recommender systems. 2017: 297-305.

[6] CHEN C, ZHANG M, LIU Y, et al. Neural attentional rating regression with review-level explanations[C]//Proceedings of the 2018 world wide web conference. 2018: 1583-1592.

[7] LIU D, LI J, DU B, et al. Daml: Dual attention mutual learning between ratings and reviews for item recommendation[C]//Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery and data mining. 2019: 344-352.

[8] SHAALAN Y, ZHANG X, CHAN J, et al. Detecting singleton spams in reviews via learning deep anomalous temporal aspect-sentiment patterns[J]. Data Mining and Knowledge Discovery, 2021, 35(2): 450-504.

[9] HUSSAIN N, MIRZA H T, HUSSAIN I, et al. Spam review detection using the linguistic and spammer behavioral methods[J]. IEEE Access, 2020, 8: 53801-53816.

[10] WANG S I, MANNING C D. Baselines and bigrams: Simple, good sentiment and topic classification[C]//Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). 2012: 90-94.

[11] SALMINEN J, KANDPAL C, KAMEL A M, et al. Creating and detecting fake reviews of online products[J]. Journal of Retailing and Consumer Services, 2022, 64: 102771.

(上接第 51 页)

[23] HUANG L, MA J Y, CHEN C L. Topic Detection from Microblogs Using T-LDA and Perplexity[C]// Proceedings of the 24th AsiaPacific Software Engineering Conference Workshops. 2017:71-77.

[24] 王晰巍, 张柳, 黄博, 等. 基于 LDA 的微博用户主题图谱构建及实证研究——以“埃航空难”为例[J]. 数据分析与知识发现,2020,4(10):47-57.