在线虚假评论检测采用监督式机器学习和 BERT 模型

[1]Abrar Qadir Mir，[2]Furqan Yaqub Khan，Mohammad Ahsan Chishti

[1,3]信息技术系，克什米尔中央大学，

查谟和克什米尔，印度，
191131
[2]计算机科学与工程系，巴特那印度理工学院，比哈尔邦，印度
801106
通讯作者：furkaan309@gmail.com

摘要

近年来，在线购物商店稳步增长。随着这类业务的大规模发展，虚假评论的检测日益受到关注。虚假评论蓄意误导消费者，从而损害在线购物环境的诚信与真实性。迄今为止，已有多种虚假评论分类器被提出，这些分类器会考虑评论的实际内容。为提高现有虚假评论分类或检测方法的准确率，我们提出使用 BERT（来自变换器的双向编码器表示）模型从文本（即评论）中提取词嵌入。词嵌入采用多种基础方法进行处理，例如支持向量机（SVM）、随机森林、朴素贝叶斯等。此外，我们还采用混淆矩阵方法来评估结果并以图形方式呈现。结果表明，支持向量机分类器的表现优于其他分类器，其准确率和 F1 分数均更高，准确率达到 87.81%，比先前研究中使用的分类器高出 7.6% [5]。

关键词： BERT 模型；机器学习；虚假评论；插入词语；控制检测

1. 引言

在线评论在各种在线购物场景中发挥着至关重要的作用，例如在线酒店预订、餐厅预订，更重要的是在电子商务网站上，评论能够帮助顾客评估不同产品的质量和价格、评价卖家的服务，并做出购买决策。如今，使用在线购物平台的用户数量持续大幅增长。以全球最受欢迎的旅游网站 TripAdvisor 为例，它约有 4.555 亿用户，拥有超过 6 亿条用户评论，涵盖超过 750 万家酒店、餐厅和航空预订服务 [1]。然而，一些唯利是图的不良商家会撰写虚假评论，误导不够谨慎的顾客，以此过度推销自己中意的产品，同时贬低其他选择。虚假评论严重影响并威胁着在线购物环境的真实性。因此，检测并移除此类虚假评论已成为当前研究中至关重要且不可或缺的领域；若处理不当，未来也可能仍是如此。虚假评论检测可以通过人工或自动方式进行，但与自动检测方法相比，人工执行这一流程被认为成本高、耗时且不准确 [2]。在

过去十年间，自动识别虚假评论方面取得了许多重大进展。许多基于机器学习的工具和技术已能有效检测虚假评论，其中一些工具和技术包括支持向量机和朴素贝叶斯[3]。本研究主要聚焦于评论内容，以检测虚假评论。我们利用自然语言处理创建一些与所提供的数据或文本没有直接关联的评论特征。我们使用 BERT（来自变换器的双向编码器表示）执行自然语言处理任务及其他文本处理工作，例如特征选择。

2. 相关研究

近年来，在线购物环境日益受到广泛关注，并在人们的生活中发挥着重要作用，既能节省时间，也能节省金钱。人们似乎相当依赖在线购物。从这一角度来看，识别在线评论同样重要，也是当今的热门研究课题之一。自 2007 [4] 以来，虚假评论检测问题就已受到关注。虚假评论检测研究主要分为两类：文本特征和行为特征。文本特征是指对评论文本进行表征，或分析并刻画评论文本的内容。行为特征则刻画评论数据语料库中的非语言特征，主要取决于评论者的行为，例如写作风格、表达情感的措辞以及评论频率。Liu [6] 提出了首个评论检测模型，以评论相似度和产品特征作为判定标准。该研究还探讨了虚假评论的质量，以及被称为“垃圾信息发布者”的用户发布虚假产品评论的行为。Ott [7] 提出了一种基于评论数据语料库文本极性的分类模型，并创建了一个包含四百条评论的数据集，将其分为真实和虚假（或人工生成）两组；随后，他使用不同的分类器并在不同的系统配置下计算模型的有效性。Savage [8] 通过加入一些新的句法特征，提高了 Ott 模型的效率。Farris [9] 提出了另一种使用遗传算法和随机权重网络对虚假评论进行分类的模型。该模型专门用于垃圾信息检测。该模型取得了出色的结果，研究据此得出结论：使用自动分类器可以凭借良好的精确率、召回率和准确率对垃圾评论进行分类。虚假评论者识别： [10] 尽管上述方法在检测虚假评论和评论内容方面效果良好，但研究者认识到，虚假评论者检测是一个额外的研究方向，仍有待开发。Huang [11] 提出了另一种观点：不真实的评论者通常会在特定时间段内频繁发表评论。因此，他为评论上传时间界定了时间范围，并估算了虚假评论的上传频率；这些频率与真实评论的上传频率存在异常差异，

即；如果上传内容达到 > 阈值；则称该评论者为虚假评论者。

Lim [12] 考察了大量评论者及相关数据，并指出虚假评论者的两个特征：1）虚假评论者会针对某个“特定”产品或某次特定的在线购物发表评论。2）真实评论者和虚假评论者在写作模式与风格以及发布评论的频率上通常有所不同。基于这两种互补的评论特征，该研究对评论者的行为进行建模，并使用另一种算法计算评论者评分。在 [13], 中，使用监督机器学习对评论进行分类。监督式分类

所采用的技术包括朴素贝叶斯、支持向量机、K-Star、K近邻和决策树。实验使用了三个不同的电影评论标注数据集。[14] 每个数据集分别包含 1400、2000 和 10662 条评论。在 [15], 中，使用了朴素贝叶斯、支持向量机和决策树技术，以及随机森林和最大熵分类器等分类器。该数据集由三星产品和服务的评论构成，其中包含 10000 条虚假评论。在 [16], 中，使用了包含 1600 条评论的芝加哥酒店评论数据集。在 [17], 循环神经网络、平均广义回归神经网络、广义回归神经网络、卷积神经网络和双向平均广义回归神经网络的被用于对数据语料进行分类。所用数据集来自 [18] 其中包含关于酒店、餐馆和医生的虚假与真实评论。分析仅考虑文本特征，未考虑行为特征。

3. 背景

对于需要准确处理、且常规算法无法胜任的时效性任务，机器学习是解析海量高速数据的首选工具。机器学习的优势在于能够根据数据模式自主学习并开发算法，还能在使用过程中利用后续数据不断改进算法[19]。机器学习算法根据标签和行动—奖励理论进行分类[20]；按照标签，分为监督式、半监督式和无监督机器学习技术；按照行动—奖励理论，则有强化机器学习技术。监督式机器学习技术需要标签和数据[21]，而无监督机器学习只提供数据，并根据所需属性或函数找出不同数据点之间的关系。半监督技术是监督式和无监督机器学习技术的混合形式。最后，强化学习会针对特定情境或行动作出反应；如果反应方向正确，就提高所采取步骤的优先级，也就是给予奖励；如果方向不正确，就降低该步骤的优先级，也就是施以惩罚。本文主要采用监督式机器学习分类器。文中使用的几种重要监督式分类器包括支持向量机，它通过超平面将所需的两个类别分开[22]；另一种是朴素贝叶斯分类器，它使用贝叶斯定理计算特定评论属于给定类别之一的概率。贝叶斯定理的公式为 P(A—B) = P(B—A)*P(A) P(B) [23]。

K近邻（KNN）算法[24] 主要用于统计估计和模式识别。距离函数用于将距离最近的属性归入一个簇，从而尽量缩小簇内距离并扩大簇间距离，以便更准确地进行分类。[25] 决策树[26] 根据基尼比率、熵、信息增益等数值对数据进行分类。因此，每个决策点都表示为树的内部节点，而每个类别则表示为树的叶节点，使该算法的表示呈现树状结构，因此称为决策树。随机森林[27]是一种集成算法，由多棵决策树组成，这些决策树通常各不相同，从而消除决策树算法中常见的过拟合问题。逻辑回归[28] 也基于逻辑函数或对数函数，在两种不同的数据类型之间构建超平面。

4. 数据文件

本文主要使用四个数据集进行实验（表1），分别是酒店、医生、餐厅和亚马逊数据集。康奈尔大学的正面评论[29] 和负面评论[30] 数据集合并后构成酒店数据集，具体如[29-30]所述。这些数据集被视为金标准模拟评论语料库[30]。土耳其人或受委托的虚假评论者负责撰写虚假评论，每位土耳其工人撰写一条评论。经过筛选，删除了简短评论和抄袭评论。研究人员要求土耳其工人撰写看起来真实可信的评论。与此相对应，真实评论取自Expedia、TripAdvisor和Hotels.com等多个在线评论网站。该数据集包含20家酒店的评论，每家酒店有80条评论，平均分为两类，共计1600条；其中一半为真实评论，另一半为虚假评论。每条评论还附有相应的真实与虚假标签。因此，数据集中的每条评论都包含真实/虚假标签、酒店信息、旅行社名称、情感极性（正面/负面）和评论内容，平均每条评论有152个词。同样，还创建了另外两个数据集[31] ，即餐厅评论数据集和医生评论数据集。研究人员为芝加哥最知名的10家餐厅各创建了20条虚假评论；医生评论数据集中的356条正面虚假评论由评分较高的美国土耳其工人撰写，并从顾客对这些酒店和医生的评论中选取了数量相同的真实评论。亚马逊的数据集包含21,000条评论，其中10,500条被亚马逊认定为虚假评论。数据还提供了评论的一些其他信息，例如类别标签、评分、是否为已验证购买、产品类别和产品编号。评论平均评分为4.13/5，且55.7%的数据来自已验证购买。评论涉及30个产品类别，每个类别有700条评论。亚马逊认定这些类别不符合其政策。

表 1 本研究使用的虚假评论数据集

数据集	虚假/真实评论	情感极性	平均评论长度（词数）
酒店 [29, 30]	400/400	正面和负面	151.9
餐厅 [31]	200/200	阳性	137.1
医生[32]	356/200	阳性	102.4

5. 所提系统

图 1 展示了所提出的方法。该方法分为 3 个阶段，并为我们提供了用于虚假评论分类的最佳模型，具体如下：

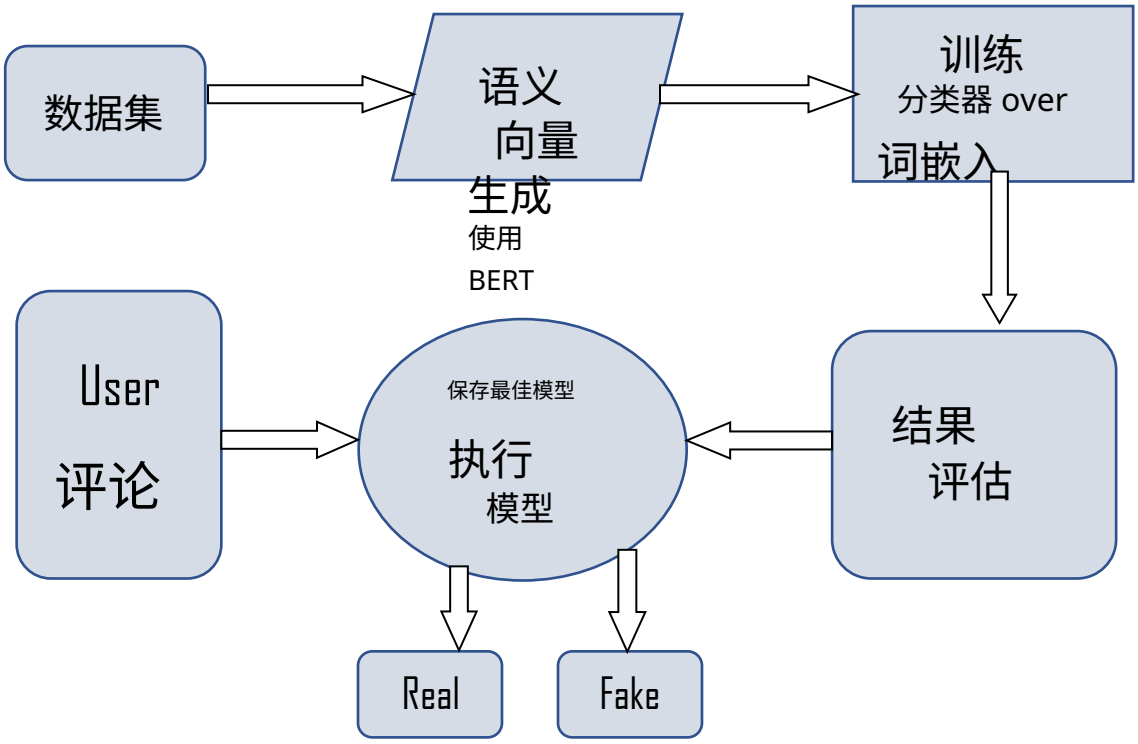


图 1：所提出模型的工作原理

1) 加载数据集 编码表示来自变换器输入到 BERT 词嵌入，即数据集中各文本词语的大型向量表示。2）随后将输入加载到分类模型中进行训练。训练数据和测试数据按 80:20 的比例划分，即 80% 用于训练，其余用于测试模型。3）使用混淆矩阵表示法评估精确率、召回率、F1分数和准确率。4）随后保存表现最佳的分类器，并在之后用于检测用户评论是虚假还是真实。

6. BERT 模型的运行原理

基于变换器的 BERT（来自变换器的双向表示）是一种注意力机制，能够学习文本中词语之间的上下文联系。简单的变换包含一个用于读取文本输入的编码器，以及一个用于生成任务预测的解码器。由于 BERT 的目标是创建语言表示模型，因此只需要编码器的一部分。输入到 BERT 编码器的是一组词元，它们首先被转换为向量，然后在神经网络中进行处理。然而，在开始处理之前，BERT 要求为输入添加一些额外的元数据。变换器本质上由一个将序列映射到序列的层构成，因此输出也是一个向量序列。

6.1 生成 BERT嵌入

使用 BERT 生成大型句子嵌入。BERT 是一种预训练模型。将句子 Si 切分为词 W = {w1, w2, w3,..,wn}。每个 w∈ W 都输入到

BERT 中，以获取 w 的词嵌入$_s$ ∈W。设 w1 的嵌入为 E1。同理，我们获得 Si 中所有 wn 个词的嵌入，因此，句子 Si 的 n 个词共有 Vn 个嵌入。然后，将 Si 定理的所有嵌入合并，形成 Si 定理的大型嵌入（BV）。因此 BE = {E1 ∪ E1 ∪ E1 ∪ ... ∪ En }。为酒店数据集生成嵌入。获得所有评论的大型嵌入及其对应标签后，我们将这些嵌入加载到分类模型中进行训练和测试，训练数据与测试数据的比例为 80:20。

7. 实验评估：

本节介绍六项实验的结果，并使用六种不同的机器学习分类器对其进行评估，分别为：支持向量机、随机森林、装袋分类器、AdaBoost 分类器、朴素贝叶斯和 K近邻分类器。我们在酒店、餐厅和医生数据集上测试了分类模型。但受篇幅限制，我们仅展示在酒店数据集上进行的实验结果。所有已实现分类器的混淆矩阵如下：

Accuracy =87.812500%

	precision	recall	f1-score	support
deceptive	0.87	0.90	0.89	170
truthful	0.88	0.85	0.87	150
accuracy			0.88	320
macro avg	0.88	0.88	0.88	320
weighted avg	0.88	0.88	0.88	320

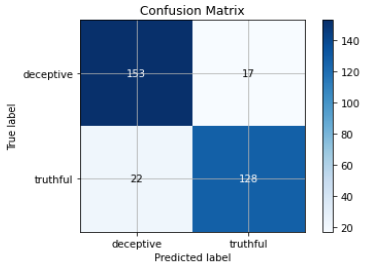


图 2：支持向量机的混淆矩阵

Accuracy =83.437500%

	precision	recall	f1-score	support
deceptive	0.85	0.84	0.84	170
truthful	0.82	0.83	0.83	150
accuracy			0.83	320
macro avg	0.83	0.83	0.83	320
weighted avg	0.83	0.83	0.83	320

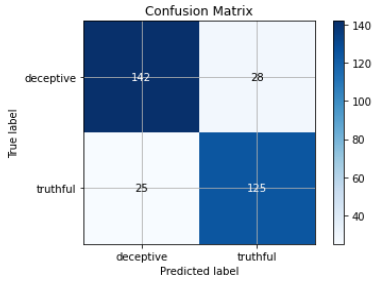


图 3：随机森林的混淆矩阵

Accuracy =79.062500%

	precision	recall	f1-score	support
deceptive	0.79	0.82	0.81	170
truthful	0.79	0.75	0.77	150
accuracy			0.79	320
macro avg	0.79	0.79	0.79	320
weighted avg	0.79	0.79	0.79	320

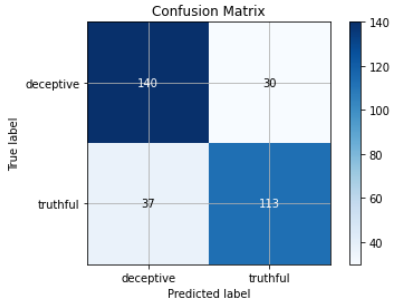


图 4：装袋分类器的混淆矩阵

Accuracy =78.437500%

	precision	recall	f1-score	support
deceptive	0.80	0.79	0.80	170
truthful	0.76	0.78	0.77	150
accuracy			0.78	320
macro avg	0.78	0.78	0.78	320
weighted avg	0.78	0.78	0.78	320

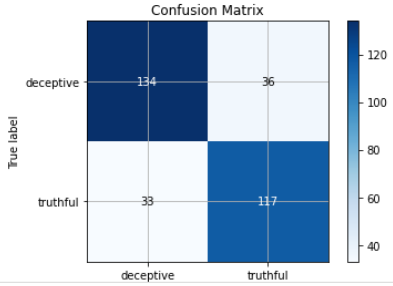


图 5：自适应提升的混淆矩阵

Accuracy =78.437500%

	precision	recall	f1-score	support
deceptive	0.81	0.77	0.79	170
truthful	0.75	0.80	0.78	150
accuracy			0.78	320
macro avg	0.78	0.79	0.78	320
weighted avg	0.79	0.78	0.78	320

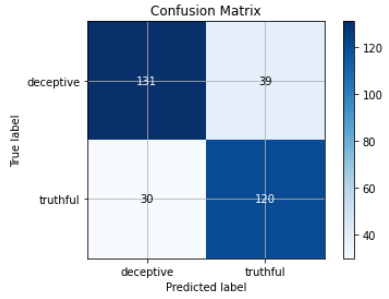


图6：朴素贝叶斯的混淆矩阵

Accuracy =77.187500%

	precision	recall	f1-score	support
deceptive	0.75	0.86	0.80	170
truthful	0.81	0.67	0.73	150
accuracy			0.77	320
macro avg	0.78	0.77	0.77	320
weighted avg	0.78	0.77	0.77	320

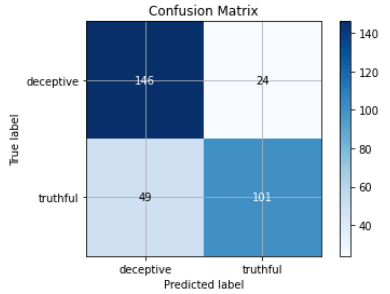


图7：K近邻分类器的混淆矩阵

分类器	准确率（%）		F分数	
	以往工作**[18]**	本研究	以往工作**[18]**	本研究
SVM	80.75	**87.81**	0.80	**0.88**
随机森林	79.31	**83.43**	0.79	**0.83**
装袋法	78.19	**79.06**	0.78	**0.79**
K-NN	71.38	**77.18**	0.68	**0.78**
自适应提升算法	77.06	**78.43**	0.77	**0.78**
高斯朴素贝叶斯	81.25	78.43	0.81	0.78

表2：所提出模型的性能比较分析

7. 结论与未来工作

本文展示了评论的重要性，以及评论如何影响网络数据相关的几乎所有方面。显然，评论在人们的决策过程中发挥着至关重要的作用。因此，虚假评论检测是一个活跃且持续发展的研究领域。本文提出了一种利用机器学习检测虚假评论的方法，并在该方法中考虑了评论属性。本文使用 Hotels 数据集对所提出的方法进行了实验评估，并在该数据集上使用了不同的分类器。结果表明，在检测虚假评论的过程中，支持向量机分类器优于其他分类器（准确率为 87.81%），因此，只考虑评论的文本内容，而不一定考虑其情感特征，就能有效地将评论分类为真实或虚假。然而，未来的工作可以考虑纳入评论者的行为特征，例如取决于评论者发表评论次数的特征，以及评论者完成评论所需的时间

评论，以及他们提交正面或负面评论的频率。考虑到行为特征有望显著提升所提出的虚假评论检测方法的性能。此外，使用神经网络模型执行此任务，同样有助于从大型数据集中检测虚假评论。

参考文献

1. TripAdvisor 首页。http://ir.tripadvisor.com/。访问日期：2019年1月21日。
2. Harris, C.：利用人类计算检测欺骗性观点垃圾信息。载于：AAAI人工智能研讨会，页码87–93。AAAI（2012年）3. Heydari, A., ali Tavakoli, M., Salim, N., Heydari, Z.：评论垃圾信息检测：综述。《专家系统应用》42(7), 3634–3642（2015年）4. N. Jindal and B. Liu，“评论垃圾信息检测”，载于《第16届万维网国际会议论文集》，系列编号WWW ’07，2007年。5. Hajek, P., Barushka, A. and Munk, M., 2020. 使用整合词嵌入和情感挖掘的深度神经网络检测虚假消费者评论。《神经计算与应用》，32(23)，页码17259-17274。6. Jindal N, Liu B（2007）分析和检测评论垃圾信息。载于：第7届IEEE国际数据挖掘会议，ICDM 2007，IEEE，页码547–552。https://doi.org/10.1109/icdm.2007.687. Ott M, Choi Y, Cardie C（2011）凭想象力的任意发挥来发现欺骗性观点垃圾信息。《计算语言学协会第49届年会论文集》6(8):309–3198. Savage D, Zhang X, Chou P（2015）基于异常评分偏差检测观点垃圾信息。《专家系统应用国际期刊》42(22):8650–86579. Faris H, Ala’m AZ, Heidari AA, Aljarah I, Mafarja M, Hassonah MA, Fujita H（2019）一种基于进化随机权重网络的智能系统，用于垃圾信息检测及最相关特征识别。《信息融合》4:67–8310. Fang, Y., Wang, H., Zhao, L., Yu, F. and Wang, C., 2020. 基于动态知识图谱的虚假评论检测。《应用智能》，50(12)，页码4281-4295。11. Huang J, Qian T, He G（2013）检测专业垃圾评论者，高级数据挖掘与应用。Springer，柏林，页码288 29912. Lim EP, Nguyen VA, Jindal N, Liu B, Lauw HW（2010）利用评分行为检测产品评论垃圾信息发布者。ACM国际知识管理会议 7(12):939–94813. E. Elmurngi and A. Gherbi，《利用机器学习技术通过情感分析检测虚假评论》。IARIA/数据分析，2017年。14. V. Singh, R. Piryani, A. Uddin, and P. Waila，“电影评论和博客文章的情感分析”，载于《先进计算会议》（IACC），2013年，页码893–898。15. A. Molla, Y. Biadgie, and K.-A. Sohn，“从推文中检测负面欺骗性观点”，载于《国际移动与无线技术会议》。新加坡：Springer，2017年。16. S. Shojaee et al.，“利用词汇和句法特征检测欺骗性评论。”2013年。

17. Y. Ren and D. Ji，“用于欺骗性观点垃圾信息检测的神经网络：一项实证研究”，《信息科学》，第385卷，页码213–224，2017年。18. H. Li et al.，“通过集体正例-未标记学习发现虚假评论。”2014年。19. D. Michie, D. J. Spiegelhalter, C. Taylor et al.，“机器学习”，《神经网络与统计分类》，第13卷，1994年。20. T. O. Ayodele，“机器学习算法的类型”，载于《机器学习新进展》。InTech，2010年。21. F. Sebastiani，“机器学习在自动文本分类中的应用”，《ACM计算调查》（CSUR），第34卷，第1期，页码1–47，2002年。22. T. Joachims，“利用支持向量机进行文本分类：基于大量相关特征的学习。”1998年。23. T. R. Patil and S. S. Sherekar，“朴素贝叶斯和J48分类算法在数据分类中的性能分析”，页码256–261，2013年。24. M.-L. Zhang and Z.-H. Zhou，“ML-KNN：一种多标签学习的惰性学习方法”，《模式识别》，第40卷，第7期，页码2038–2048，2007年。25. N. Suguna and K. Thanushkodi，“利用遗传算法改进K近邻分类”，《国际计算机科学问题杂志》，第7卷，第2期，页码18–21，2010年。26. M. A. Friedl and C. E. Brodley，“基于遥感数据的土地覆盖决策树分类”，《环境遥感》，第61卷，第3期，页码399–409，1997年。27. A. Liaw, M. Wiener et al.，“随机森林的分类与回归”，《R新闻》，第2卷，第3期，页码18–22，2002年。28. D. G. Kleinbaum, K. Dietz, M. Gail, M. Klein, and M. Klein，《逻辑回归》。Springer，2002年。29. Ott M, Cardie C, Hancock J（2012）估算在线评论社区中欺骗行为的流行程度。载于：第21届万维网国际会议，ACM，页码201–210。https://doi. org/10.1145/2187836.2187864 30. Ott M, Cardie C, Hancock JT（2013）负向欺骗性观点垃圾信息。载于：2013年北美计算语言学协会分会会议：人类语言技术，ACL，页码497–501 31. Li J, Ott M, Cardie C, Hovy E（2014）探索识别欺骗性观点垃圾信息的一般规则。载于：计算语言学协会第52届年会论文集，ACL，第1卷，页码1566–1576。32. Garcia L（2018）亚马逊上的欺骗——一次自然语言处理探索。https://medium.com/@lievgarcia/deception-on-amazon- c1e30d977cfd。访问日期：2019年9月1日