使用监督学习识别虚假评论的数据分析

阿尔萨德·奥萨玛，泰亚兹·阿卜杜拉
（所属单位：印度计算机科学与信息技术系）

摘 要：虚假评论，也称为欺骗性意见，用于误导人们，最近变得越来越普遍。这是由于在线营销交易的快速增长，例如销售和采购。电子商务为客户提供了一种设施，可以在购买时发布有关产品或服务的评论和评论。新客户通常会在做出购买决定之前浏览网站上发布的评论或评论。然而，当前的挑战是新人如何区分真实评论和虚假评论，从而欺骗客户、造成损失并损害公司声誉。本文尝试开发一种智能系统，该系统可以使用评论文本的 n-gram 和评论者给出的情感分数来检测电子商务平台上的虚假评论。本研究采用的拟议方法使用标准的酒店评论数据集进行实验和数据预处理方法，并使用词频 - 逆文档频率 (TF-IDF) 方法提取特征。对于检测和分类，本文将评论文本的 n-gram 输入到构建的模型中，以分类为虚假或真实。然而，实验是使用四种不同的监督机器学习技术进行的，并在从 Trip Advisor 网站收集的数据集上进行了训练和测试。这些实验的分类结果表明，朴素贝叶斯 (NB)、支持向量机 (SVM)、自适应增强 (AB) 和随机森林 (RF) 分别获得了 88%、93%、94% 和 95% 的基于 关于测试准确性和 tje F1 分数。作者将获得的结果与使用相同数据集的现有工作进行比较，所提出的方法在准确性方面优于可比较的方法。
关键词：电子商务；虚假评论检测；方法；机器学习；酒店点评

Data Analytics for the Identification of Fake Reviews Using Supervised Learning

Alsaade Osamah, Theyazn Abdullah
(Affiliation: Department of Computer Science & Information Technology, India)

Abstract: Fake reviews, also known as deceptive opinions, are used to mislead people and have gained more importance recently. This is due to the rapid increase in online marketing transactions, such as selling and purchasing. E-commerce provides a facility for customers to post reviews and comment about the product or service when purchased. New customers usually go through the posted reviews or comments on the website before making a purchase decision. However, the current challenge is how new individuals can distinguish truthful reviews from fake ones, which later deceives customers, inflicts losses, and tarnishes the reputation of companies. The present paper attempts to develop an intelligent system that can detect fake reviews on ecommerce platforms using n-grams of the review text and sentiment scores given by the reviewer. The proposed methodology adopted in this study used a standard fake hotel review dataset for experimenting and data preprocessing methods and a term frequency-Inverse document frequency (TF-IDF) approach for extracting features and their representation. For detection and classification, n-grams of review texts were inputted into the constructed models to be classified as fake or truthful. However, the experiments were carried out using four different supervised machine-learning techniques and were trained and tested on a dataset collected from the Trip Advisor website. The classification results of these experiments showed that naïve Bayes (NB), support vector machine (SVM), adaptive boosting (AB), and random forest (RF) received 88%, 93%, 94%, and 95%, respectively, based on testing accuracy and tje F1-score. The obtained results were compared with existing works that used the same dataset, and the proposed methods outperformed the comparable methods in terms of accuracy.
Keywords:E-commerce; fake reviews detection; methodologies; machine learning; hotel reviews

引言

随着电子商务的快速发展，产品和服务的在线采购和销售成为一种趋势，客户越来越多地使用这个在线营销网站进行采购以满足他们的需求。购买后，客户会写下关于他们对产品和服务的个人经历、感受和情绪的评论。在线产品评论在电子商务业务的成败中起着重要作用。购物者在购买产品或服务前，通常会先浏览过往顾客在网上发表的评论，以了解有关产品详情的建议，从而作出购买决定。然而，可以通过发布虚假评论来增强或阻碍特定的电子商务产品，这些评论可以由被称为欺

诈者的人撰写。 这些评论可能会给电子商务企业造成财务损失，并误导消费者做出错误的决定来搜索替代产品。

大部分正面评价会吸引更多客户购买特定产品或品牌。 正面意见提供可观的经济收益，而负面意见通常会导致电子商务的销售损失。 因此，大多数商家主要依靠公众舆论来通过提高产品质量来重塑他们的商业计划。通常，意见是任何在线博客、帖子或评论的关键。 垃圾内容可以定义为合并成意见并用于广告、促销、传播信息和经济利益目的的无意义或未经请求的数据。 消费者为获得产品或服务做出在线购买决定，为此，他们会在购买前查看电子商务网站上提供的在线产品评论。 虚假评论检测系统是自然语言处理的一个子领域。 它旨在分析、检测和过滤评论者的评论，特别是在电子商务网站上，将其转化为虚假或真实的评论。 虚假意见是指评论中的虚假或不准确信息，误导消费者做出错误的购买决定，影响商品收益。 垃圾意见可分为三种类型： 1) 故意编写的不真实（虚假）评论，以误导读者或意见挖掘系统。 它们包括对特定目标产品的不值得的正面评价，以推广产品或服务。 此外，它们包括对有价值产品的负面评论以诽谤它们。这些被命名为类型 1 垃圾邮件内容。2) 仅品牌评论的特点是针对品牌而非产品本身的主观意见。 这样的评论被称为类型 2 垃圾评论。 3) 非评论有两种子类型：(a) 广告和 (b) 不包含意见的不相关评论，例如问题、答案或不明确的文本。

根据意见垃圾邮件检测领域的文献，没有特定的特征来区分真假内容。 因此，本研究旨在通过监督学习算法提高虚假评论识别系统的准确性。 为此，从评论文本中提取特征是一项重要且有意义的任务。 这些特征是情感得分、强阳性词、强阴性词和四克，以及动词、名词和形容词的数量。

用于虚假 / 垃圾邮件评论和垃圾邮件发送者检测的数据集和技术

自 Jindal 首次提出“意见垃圾邮件”的概念以来，人们已经进行了研究。 总体而言，垃圾评论（意见垃圾邮件）检测方法可分为两种类型：监督学习和无监督方法。当提供足够的标记数据时，通常使用监督学习方法。给定正确的特征和标记的训练数据，它具有相对较好的性能。监督学习的优点是可以利用标记数据的潜在特征，利用先验知识进行分类任务。 该方法的缺点是它需要相当数量的标记数据，这可能需要更多的劳动来标记，特别是当需要人工标记时。 Jindal 的研究建立了一个分类器，利用重复或接近重复的评论作为虚假评论，其余的作为非虚假评论。 Ott 和 Law 的作者利用评论内容词性（POS）、LIWC 文本特征和语言模型来发现欺骗性意见垃圾邮件，但没有考虑用户行为这对检测虚假信息非常有用 审稿人和审稿人。 无监督方法已用于检测群组垃圾邮件发送者并审查突发性和行为足迹。

Mukherjee 提出了一种寻找候选评论者群体并根据群体之间的关系建立关系模型的方法。 我们的工作仍然使用“组”的概念，但考虑的是评论之间的关系，而不是评论者，这与他们的工作截然不同。 费等人利用评论的突发性来识别评论垃圾邮件发送者。戈斯瓦米等人提出了一项关于评论者社交互动对在线消费者评论中欺诈检测的影响的研究。 在他们的实验中，收集并预处理了 Yelp 的评论数据集（135,413 条评论，其中 103,020 条是推荐评论，32,393 条是不推荐评论）。 然后，提取用户的行为和社交互动特征，并采用反向传播神经网络算法对评论进行真实和欺诈分类。

金达尔等人报告了垃圾邮件 / 虚假评论检测的第一项研究。 作者识别出三种类型的垃圾评论，它们是不真实的、仅针对品牌的评论和不相关的评论。 他们使用监督机器学习技术（逻辑回归）将亚马逊产品评论的重复和接近重复分类为垃圾邮件或非垃圾邮件，得到的结果在曲线下面积（AUC）方面为 78%。 穆克吉等人提出了一种 SVM 模型来检测虚假产品评论。 在数据集方面，他们使用了真实的 Yelp 产品评论，其中包括来自酒店的 5,678 条评论和 5,124 名评论者，以及来自餐厅的 58,517 条评论和 35,593 名评论者。在他们的实验中，提取了两种类型的特征：语言特征，包括 n-gram、词性和 LIWC，以及评论者的行为特征。 为了确定虚假和非虚假（真实）评论词的两种分布之间的差异，他们应用了 Kullback – Leibler 散度 (KL)。 他们的方法的结果是使用语言特征的准确率为 84%，使用审阅者特征的准确率为 86%。

艾哈迈德等人提出了一种线性支持向量机技术，用于基于 N-gram 特征的虚假评论检测。 作者评估了从 Tripadvisor.com 收集的标准虚假酒店评论数据集。 对于特征提取，使用了 TF-IDF 方法，准确率达到 90%。 李等人试图定义识别欺骗性评论的一般规则。 在他们的方法中，跨域数据集包括从 Amazon Mechanical Turk 收集的 800 条酒店评论，以及 400 条来自领域专家的欺骗性医生评论。 在特征方面，他们的方法使用了 unigram、LIWC 和 POS。 他们通过使用稀疏加性生成模型 (SAGE) 来使用多类分类方法，该模型由广义加性模型和主题模型的组合组成。此外，SVM 应用于相同的数据集和特征。本次实验取得的准确率分别为 81% 和 78%。

方法

在本实验中，黄金标准数据集由 Ott 等人开发。它包含从一个流行的酒店预订网站 Trip Advisor 收集的 1,600 条酒店评论。 该数据集的作者对来自芝加哥 20 家酒店的所有 5 星级和 3 星级评价进行了提炼。 通过添加评论长度、四克、情感分数和 POS 等特征对数据集进行预处理。

审查文本

文本格式的用户写评论内容。 基于此特征，主要分析任务用于获得文本特征，例如情感分析和语言特征。

情感评分

情感分数是用于计算和查找给定文本的极性分数（正面、负面或中性）的过程。 负面欺诈者通常习惯于在他们的评论中包含负面词语，而不是正面词语，当他们表现出明显的负面情绪时。 同理，正面的骗子总是习惯写比较肯定的词； 因此，应该为每个评论文本计算情感分数。 以下公式用于查找数据集中每个评论文本的情感分数： S (r) = P(W) − N (W) T (W) (1) 其中 S (r) 表示情感 (S) 评价 ; P(W) 指的是正词的个数； N(W) 表示否定词的个数； T(W) 表示评论文本中正面和负面词的总数。

审核时长

在本节中，使用了 POS 功能。 词性标注是根据文本内容中的每个词在句子中的位置和上下文为其附加词性标签的过程。基于此方法，从评论文本中提取形容词、名词、介词、并列连词、限定代词、动词、预定词和副词的数量。 这部分的结论是，真实评论有更多的名词和形容词，而虚假评论有更多的动词和副词。

N-Grams

从文本内容中选择 N 个相邻词作为特征的过程称为 N-gram 特征。 当一次分配 N = 1（一个词）时，它被称为 unigram。 如果一次选择两个相邻的词（N = 2），则称为二元组，同理，当同时分配四个相邻的词（N = 4）时，称为 四克。

预处理步骤

在执行特征提取步骤之前，数据需要进行某些清洗，例如标点符号去除以从评论文本中去除标点符号（？！：;，，“。），停用词去除以从文章单词中清洗评论句子（'the"a"an"'in'），从整个数据集中剥离不需要的词和字符，并进行数据分词，将评论内容的每句话拆分成单独的词、关键词、短语和信息片段。

特征提取（TF-IDF）

TF-IDF 指的是词频 – 逆文档频率； 它被认为是文本分类系统中使用的特征提取和表示方法之一。 它用于自然语言理解和信息检索。 此外，TF-IDF 是一种统计方法，用于衡量术语或单词对数据集中文档的重要性。它有两部分：词频，用于计算文档中特定词的频率，以发现文档之间的相似性。

监督机器学习技术

本小节介绍用于将评论文本分类为虚假或真实的不同监督算法。 在将数据集转换为 TF-IDF 特征形式后，在开始训练机器学习分类器之前，将数据集分为 80% 的训练集和 20% 的测试集。 在这个实验中，应用了四种不同的监督分类器，即支持向量机 (SVM)、朴素贝叶斯 (NB)、随机森林 (RF) 和自适应增强 (AB)。

支持向量机

SVM 是一种流行的监督概率算法，可用于按顺序和非顺序划分数据。 SVM 用于文本分类并在高维向量空间中提供良好的效率。 此外，它表示空间地图中的数据训练样本。 各种类别的数据点由超平面内的最大边距区分。 它的决策边界是解决训练样本的极限边界。 该方法应用了径向基函数（RBF）。

朴素贝叶斯

朴素贝叶斯 (NB) 是一种用于分类的监督式机器学习方法。 给定先前发生的另一事件的可能性，它可用于计算事件发生的可能性。 它基于条件概率定理。 通过文本分类任务，数据包含高维，这意味着每个单词代表数据中的一个特征。 但是，该模型预测文本句子中每个单词的概率，并将其视为任何一个数据集类别的特征。

随机森林

随机森林 (RF) 是机器学习技术中广泛使用的方法。RF，顾名思义，就是一片树木的森林。 它由几个决策树组成，可以帮助做出决定。 随机森林中的每棵树都是通过制作单个决策树的相同策略制作的。通过做出决定，将获得一个小决策树的投票，并且一个班级将由多数票决定。 RF 被称为分而治之的方法。 它使用一些弱学习器来生成强线性关系。 分类器中的每棵树都有一个由 N 个数据点或样本构成的根节点。 树中的每个节点 t 还包含位于分裂 St 的 Nt 个数据点，用于创建两个子节点，即 tL（左节点）和 tR（右节点）。

自适应升压

自适应提升是一种监督机器学习技术，它与学习提升分类器的特定方法有关。 它是一种分类方法，用于从弱学习器的线性组合中构建强学习器。 在自适应 boost 模型中，每个训练样本利用一个权重来决定被选入训练集的概率，最后的分类投票是根据弱学习器的加权投票进行的。

实验结果与讨论

本小节介绍了根据标准虚假酒店评论数据集评估四种不同监督分类器效率的实验结果。 在特征方面，使用 POS 和四克以及情感分数来训练和测试提出的分类器，即 NB、RF、Ada Boost 和 SVM。 拟议分类器的主要任务是检测评论文本并将其分类为虚假评论或真实评论。通过比较所提出的分类器的分类结果，RF 分类器在检测虚假评论方面提供了最佳性能，并且以 95% 的准确度和 F1 分数指标优于其他分类器。 通过 RF 进行的样本分类基于多决策树的多数表决。 Adaboost 分类器提供了相同数量的正样本和负样本，并且具有比 SVM 和 NB 分类器更好的结果，具有 94% 的灵敏度指标。 朴素贝叶斯分类器的误分类率最高，产生了 88% 的准确率和 F1 分数指标。

结论

虚假评论会影响客户和电子商务领域。 因此，虚假

评论识别在学术研究和商业领域引起了极大的兴趣。基于虚假酒店评论，研究并实现了四种监督机器学习技术，即朴素贝叶斯、支持向量机、随机森林和自适应提升，用于虚假评论识别。对于特征提取，使用了 TF-IDF 方法。通过比较实验的分类结果，随机森林分类器在检测虚假评论方面提供了更好的性能，并且优于其他分类器，达到了 95% 的准确率和 F1 分数指标。Adaboost 分类器获得了更高 (94%) 的灵敏度指标。作者对虚假评论检测的方法进行了比较分析，其中包括特征提取方法以及使用的数据集。根据目前的研究，大多数介绍的研究都使用了相同的特征提取方法。查阅文献后，没有发现大型标记的虚假评论数据集。许多研究人员使用了 Narayan 等人创建的小型数据集。然而，实验结果表明，所提出的模型优于比较方法。

参考文献

[1] S. N. Alsubari, S. N. Deshmukh, M. H. Al-Adhaileh, F. W. Alsaade and T. H. Aldhyani. Development of integrated neural network model for identification of fake reviews in e-commerce using multidomain datasets. Applied Bionics and Biomechanics, 2021;2021:1 - 11.

[2] Y. Li, X. Feng and S. Zhang. Detecting fake reviews utilizing semantic and emotion model. 2016 3rd Int. Conf. on Information Science and Control Engineering, Beijing, China, 2016;317 - 320.

[3] X. Hu, J. Tang, H. Gao and H. Liu. Social spammer detection with sentiment information. 2014 IEEE Int. Conf. on Data Mining, Shenzhen, China, 2014;180 - 189.

[4] F. Long, K. Zhou and W. Ou. Sentiment analysis of text based on bidirectional LSTM with multihead attention. IEEE Access, 2019;7:141960 - 141969.

[5] V. W. Feng and G. Hirst. Detecting deceptive opinions with profile compatibility. Proc. of the 6th Int. Joint Conf. on Natural Language Processing, Nagoya, Japan, 2013;14 - 18.

[6] S. J. Delany, M. Buckley and D. Greene. SMS spam filtering: Methods and data. Expert Systems with Applications, 2012;39(10):9899 - 9908.

[7] L. Li, B. Qin, B. W. Ren and T. Liu. Document representation and feature combination for deceptive spam review detection. Neurocomputing, 2016;254:33 - 41.

[8]S. Sarika, M. S. Nalawade1 and S. S. Pawar. A survey on detection of shill reviews by measuring its linguistic features. Int. J. Emerg. Trends Technol. Comput. Sci, 2014;3(6):269 - 272.

[9] Q. Peng. Store review spammer detection based on review relationship. Advances in Conceptual Modeling. Berlin, Heidelberg, Germany: Springer, 2014;287 - 298.

[10] N. Hussain, H. T. Mirza, I. Hussain, F. Iqbal and I. Memon. Spam review detection using the linguistic and spammer behavioral methods. IEEE Access, 2020;8:53801 - 53816.