分类号_____ 密级 _____
U D C _____ 编号 10741

硕 士 学 位 论 文

(专业学位)

论文题目 基于 PU 学习算法的网购虚假评论识别应用

研 究 生 姓 名：杨梦玲

指导教师姓名、职称：黄恒君 教授

学 科、专 业 名 称：统计学 应用统计硕士

研 究 方 向：大数据分析

提 交 日 期：2020 年 6 月 8 日

独创性声明

本人声明所呈交的论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

学位论文作者签名：杨梦论 签字日期：2020.6.8

导师签名：[illegible] 签字日期：2020.6.8

导师(校外)签名：郭立平 签字日期：2020年6月8日

关于论文使用授权的说明

本人完全了解学校关于保留、使用学位论文的各项规定，同意（选择“同意”/“不同意”）以下事项：

1.学校有权保留本论文的复印件和磁盘，允许论文被查阅和借阅，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文；

2.学校有权将本人的学位论文提交至清华大学“中国学术期刊（光盘版）电子杂志社”用于出版和编入 CNKI《中国知识资源总库》或其他同类数据库，传播本学位论文的全部或部分内容。

学位论文作者签名：杨梦论 签字日期：2020.6.8

导师签名：[illegible] 签字日期：2020.6.8

导师(校外)签名：郭立平 签字日期：2020年6月8日

Identification and Application of false comments in online Shopping based on PU Learning algorithm

Candidate : Yang Mengling

Supervisor: Huang Hengjun

摘 要

电子商务的日益发展，改变了人们的日常消费习惯，网上购物成为消费主要途径。在线评论作为消费者购物的一个重要依据，成为商家和买家关注的焦点。好评率高的商家店铺更容易获得消费者的青睐。为提升店铺好评率，部分商家利用虚假评论误导消费者来获取利益。监管部门对于虚假评论店铺也制定了惩罚措施，并对虚假评论进行识别，但是商家进行虚假评论的方式更加隐蔽，很难利用人工方法识别海量评论信息。

为快速准确地识别虚假评论，本文试图建立一套虚假评论识别体系，包括：数据源获取、文本数据清洗、训练集标注、模型选择与模型应用。

首先通过专家指导和机器学习标注相结合构建训练数据集，降低真实评论错误标注的比例，从而提高训练数据的预测能力。其次使用半监督学习的方法，利用少量标记样本，减少标记样本的工作量，利用 PU 学习算法与朴素贝叶斯、支持向量机、fastText、GBDT、XGBoost、LightGBM 不同分类器进行训练，选取最优分类器与 PU 学习算法结合。最后对预测结果进行可视化分析，对比虚假评论和真实评论之间的差异。PU 学习算法是一种半监督学习，通过将所有正样本和未标记样本进行随机组合来创建训练集。简化了数据标注的流程并提高了分类精度。PU 学习算法尤其适用于正例的数量有限并且拥有大量未标记的数据情况，该算法在虚假评论识别领域得到广泛应用。

作为应用，利用网络爬虫技术采集电商平台的商品评论实例数据。通过专家指导和机器学习方法部分标注真实评论数据，利用 PU 学习算法进行分类。实例结果表明：本文方法具有良好的虚假评论识别的性能，这为消费者和监管部门提供了新的方法，具有实际应用价值。

关键词：虚假评论 分类器 半监督学习 PU 学习算法 网络爬虫

Abstruct

The increasing development of e-commerce has changed people's daily consumption habits, online shopping has become the main way of consumption. Online reviews, as an important basis for consumer shopping, have become the focus of attention of merchants and buyers. Merchant stores with high praise rates are more likely to win consumers' favor. In order to improve the store's positive rate, some merchants use false reviews to mislead consumers to gain benefits. Supervisory departments have also formulated punitive measures for fake review shops and identified fake reviews, but the way merchants conduct fake reviews is more concealed and it is difficult to use artificial methods to identify massive reviews.

In order to quickly and accurately identify false comments, this thesis attempts to establish a set of false comment recognition methods, including: data source acquisition, text data cleaning, training set annotation, model selection and model application.

Firstly, a training data set is constructed through the combination of expert guidance and machine learning annotation to reduce the proportion of false annotations of real reviews, thereby improving the prediction ability of the training data. Secondly, using semi-supervised learning method, using a small number of labeled samples to reduce the workload of labeled samples, using PU learning algorithms and Naive Bayes, support vector machine, fastText, GBDT,

XGBoost, LightGBM and different classifiers to train and select the optimal classifier to combine PU learning algorithms. Finally, a visual analysis of the prediction results is performed to compare the differences between fake and real reviews. The PU learning algorithm is a semi-supervised learning that creates a training set by randomly combining all positive samples and unlabeled samples. Simplified data labeling process and improved classification accuracy. The PU learning algorithm is particularly suitable for cases with a limited number of positive examples and a large amount of unlabeled data. This algorithm is widely used in the field of false comment recognition.

As an application, we use the web crawler technology to collect product review instance data of the e-commerce platform. Annotate real review data through expert guidance and machine learning methods, by using PU learning algorithms for classification. The example results show that the method in this thesis has good performance of false comment recognition, which provides a new method for consumers and regulatory authorities, and has practical application value.

Keywords: False comment; Classifier; Semi-supervised learning; PU learning algorithm; Web crawler

目 录

1 绪论

1.1 选题依据

近些年，随着互联网的普及和网络技术的不断发展，我们从 web1.0 时代进入 web2.0 时代。网络用户不仅能从网站上获取各个渠道的信息，也可以在网上发表自己的想法和见解，为网站的内容提供素材并使网站内容更加个性化。

随着 Web2.0 时代的到来和电商平台的迅速崛起，购买途径的多样性和网络购物的便捷化，使得网购用户大量增加[1]。智研咨询发布的《2018-2024 年中国电商行业市场竞争现状及未来发展趋势研究报告》[①]显示，中国电子商务市场交易规模在 2018 年预计达到 28.4 万亿，网络购物市场规模占比扩大，预估从 2010 年的占比 11.3%提升至 27.3%；目前中国电子商务市场日益扩大，网络购物的规模逐渐增加，网络零售市场交易规模增速保持稳定。

便捷的网上购物，使消费者们越来越倾向在电商平台上购买心仪的商品。商品的在线评论数量伴随着销售量的增加而增加。在线评论有着众多方面的现实意义：对于商品制造商而言，根据顾客的产品反馈与评价，有针对性地提高产品质量，进行对应的产品优化创新；对于电商零售商而言，总结分析买家的意见和建议，更好地把关商品品质，改善服务态度，进而提升市场竞争力；对于潜在客户而言，通过在线评论，可弥补网购不能亲身体验产品质量的不足，也能够对产品有深刻的认识[2]。顾客在购买商品前，不同商店在线评论的真实性、使用感受、信息反馈等，会比相应商家的商品陈述更有可信度[3]。

据研究发现，正面评论有利于吸引顾客购买某一产品，也有利于带来显著的收益；反之负面评论会导致销量下降，收益减少[4]。由于互联网的开放性，任何人都可以轻松地撰写评论并将其发布到网络上。某些制造商和零售商发现这些特性，雇佣网络刷手进行刷单，刻意地提高商品的好评数量和店铺评分。刷单行为首先诱使消费者浏览并购买自家的产品，其次造成不公平诋毁竞争对手的现象，这种行为使得在线评论存在一定的虚假性，丧失在线评论最初存在的意义和价值。

刷单行为一般分为两类，刷好评和恶意差评。刷好评的目的主要有：增加商品的销售量、提高店铺的综合排名、产生大量好评信息来吸引顾客；恶意差评主要是降低竞争

[①] 中国产业信息网 https://www.chyxx.com/research/201806/649721.html

商家的信誉。随着电商平台监管力度的加强，刷单行为也在不断变化，现有的刷单行为主要通过以下方式进行：一是商家召集身边的亲戚朋友购买商品并发表虚假好评。这种方法成本较低但成交量较少且效率低下，不能大幅提升店铺的销售量和口碑。二是商家雇佣专门的刷单公司或者团队进行的刷单行为。这种专业人员的刷单行为与真实顾客的购买流程相似，经过一系列的培训，刷单人员能降低平台身份识别次数、躲避平台核查。根据艾媒咨询[②]发布的《2015 中国电商“刷单”现象调查报告》显示，有近八成的刷单卖家中倾向于考虑通过专业的第三方平台进行刷单，现有电商平台中许多商家都存在刷单行为。

网购热潮下，刷单行为层出不穷，虚假评论有误导消费者网购、混淆真实用户评论、降低评论信息的参考价值等诸多弊端。人工识别效率低下且成本较高，需要构建一种快速有效过滤虚假评论的识别体系。如何准确有效的识别虚假评论有着深远的意义。

1.2 国内外文献综述

本章节分两部分介绍网购虚假评论方面的研究。第一部分介绍了近十年虚假评论识别方法方面的研究成果，并进一步对这些方法存在的优缺点进行分析。第二部分阐述了数据集构造方面的研究。考虑到模型的适用性，本文试图构造一种通用的数据集构造方法。

1.2.1 虚假评论识别方法相关研究

Jindal 等[5]在 2007 年首次提出了虚假评论或垃圾评论这一概念，并对虚假评论进行研究。目前，关于虚假评论的研究主要集中在三个领域：一是虚假评论形成动机，二是虚假评论对消费者购买意愿的影响，三是虚假评论识别和虚假评论防治[6]。本文主要开展虚假评论的识别工作，国内外学者在这领域已有如下研究：

从评论内容本身出发，现有研究主要通过寻找合适的分类模型并设计显著有效的特征来提高分类器的效果。Jindal 等[7]在 2008 年提出识别垃圾评论模型，将垃圾评价定义为三个类别：虚假评论(虚假好评和恶意差评)、针对品牌的评论(不会对产品本身发表评论，而是对品牌或制造商或销售商发表意见)、无关评论，将重复评论定义为虚假评论并作为机器学习模型训练的正样本，其余评论作为负样本，挖掘显著特征并使用逻辑回

② 艾媒网 https://www.iimedia.cn/c400/39696.html

归模型来识别英文虚假评论。Li 等[8]创建跨越领域的数据集(酒店、餐厅和医生)，基于稀疏可加生成模型 SAGE(the Sparse Additive Generative Model)和 SVM 算法，使用 POS(Part of speech)、LIWC(Linguistic Inquiry and Word Coun)、Unigram 等特征识别虚假评论。Li 等[9]提出将时空特征融入到 SVM 模型。通过时间数据和空间数据的结合，发现用户注册 IP 位置与虚假评论存在一定关联，准确率提升了 12%。但是现在电商平台为保护用户个人隐私，获取不到公开的注册 IP 与评论 IP 地址，该研究难以大范围推广。

虽然有监督学习进行虚假评论的识别效果较好，但是进行监督学习需要大量的标注样本，标注样本获取需要消耗大量的资源，无监督学习虽不需要标记数据但识别效果较差。

Raymond 等[10]在 2011 年创新性地提出一种将语义语言模型(SLM)与文本挖掘相结合的无监督文本挖掘模型，并应用于 Amazon 收集的实例数据集中，实验结果表明无监督的方法可以较好地识别重复虚假评论。宋海霞等[11] 提出了一种自适应聚类的虚假评论检测方法。这种方法可以提取评论自身基本特征以及与其他评论之间的关联性特征，利用F统计量对K均值聚类方法进行调整。但是在验证模型的准确性时需要人工标注，失去了无监督学习不需要标注样本的特性。

在其他领域中，学者 Zhu 等[12]研究已经发现：与全监督学习方法相比，半监督学习介于监督学习(完全标记样本)和无监督学习(完全未标记样本)之间，结合未标记与少量标记的数据可以明显提高模型预测的准确性。有众多学者发现，在识别虚假评论领域，半监督学习也能有效的识别虚假评论。

Li 等[2]提出通过协同训练(co-training)算法双视图的半监督方法识别未标记虚假评论。结果表明，双视图协同训练算法可以获得比单视图算法更好的结果。2003 年，Liu 等[13]提出了 PU 学习(Positive-unlabled Learning)，该算法是一种半监督学习方法，使用正样本(positive examples)和未标注样本(unlabeled examples)的数据进行训练学习。PU 学习是一种迭代算法，它通过在未标注数据中去识别可靠的负样本，使用所有未标记的数据作为负样本训练和评估模型并移除被分类为正样本的任何实例，重复上述过程直到满足阈值条件。2013 年，Hernández 等[14]运行PU学习进行虚假评论的识别，使用 Ott 等[15]创建的数据集，并使用 F-Measure 作为评判指标评估模型的优虏，Hernández 等使用朴素贝叶斯和 SVM 作为训练数据的分类器。选取 100 个正例进行 PU 学习得到较高的识别效果。但是由 Ott 等创建的数据集可能无法提供现实评论的特征。任亚峰等[16]利用少量的真实评论和大量的未标注评论，提出一种混合种群性和个体性的PU学习算法，即

通过将虚假评论识别问题引入到PU学习中，由于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。

虚假评论的产生往往不是一个个体的行为，而是一个团体的集体行为，因此有许多学者从虚假评论者或者刷单团体的角度进行虚假评论识别。

Lim 等[17]在 2010 年使用 Amazon 评论识别出平台中存在的发布虚假评论的用户。为检测出进行虚假评论的人，他们定义了虚假评论人具有以下比几种行为特征，如虚假评论人可能会对某一特定店铺发表多条虚假评论，并创新性地提出一种线性加权的评分方法，进行星级评分。这种检测虚假评论人所使用的方法是一种间接地判定虚假评论的方法。但是 Lim 的研究存在一定的局限性，比如模型效果会依赖用户发表的多篇评论内容，但在实际生活中，大多数用户并不会对商品进行过多的评论。2011 年 Wang 等[18]提出一种基于异构评论的图模型，这个模型没有使用评论的文本信息，而是通过获取评论人、评论内容和商家之间的关系构造迭代模型，并计算三者之间的信任度得分识别虚假评论。Li 等[9]根据评论人、评论内容和买家IP地址之间的关系构建出“用户—评论—IP”图，基于半监督学习提出 CPU (Collective Positive-Unlabeled Learning)模型来识别虚假的评论。由于目前网络信息安全的监管，很难获取评论者的 IP 地址，该方法不宜大范围推广。2015 年，富越等[19]通过伪装刷客潜入第三方刷单平台获取刷客信息，通过分析评论者的行为从评论者、商品、评论、商家四个方面提取 14 个相应特征，采用 SVM 算法和 KNN 算法构建分类模型，并使用两种模型对淘宝网上的刷客进行识别，但是不同刷单平台的刷客特征会有所不同，由此开展的识别工作会有一定的片面性。

通过查阅与学习国内外的文献资料，发现对于识别虚假评论的解决方法大致分为两个方向：一是从评论者出发，通过检测虚假评论发表者识别虚假评论；二是从评论内容本身出发，利用监督学习、无监督学习和半监督学习等方法开展识别工作。

从评论者的角度出发，通过评论者的行为或评论等相关特征开展，但是由于不同平台开放程度不同，可获取到的数据指标存在着差异，模型移植性能较差。从评论内容出发，获取评论信息相对较为方便，但监督学习只能使用标注数据作为训练集，标注大量数据需要耗费大量人力物力且无监督学习在模型验证或其他阶段也需要标注数据。因此，本文识别电商平台的虚假评论，从评论内容角度出发，选取半监督学习中 PU 学习算法，使用少量标注数据和大量未标注数据构建分类模型。

1.2.2 数据集构造相关研究

在现有研究中，学者们在虚假评论领域研究所使用的训练集主要有两大来源：商业评论数据和众包评论数据[20]。第一类是商业评论数据，来源于现实世界中的购物平台或点评网站，但是大部分为未标记数据。第二类来自众包平台，雇佣其他人员在平台上发表虚假评论，众包平台发表的评论即为虚假评论，但是该方法花费成本较高，且与现实评论相距甚远。

目前学者们研究虚假评论所用数据集一般通过下列方式：将重复评论标注为虚假评论[7]、人工标注训练集[21]、利用众包平台生成虚假评论[15]。这些方法获得的训练集中包含的虚假评论存在一定不合理之处，与现实的情况存在一定的偏差，因此构造一个标准的训练集对于识别虚假评论而言至关重要。

康奈尔大学的 Ott 等[15]利用众包平台构建一个大型的黄金标准数据集，有针对性的为某酒店撰写 400 条五星好评，再从 TripAdvisor 网站上为相同的酒店收集 400 条真实的五星评价，该数据集包含对酒店有着积极倾向人工制造的虚假评论和真实评论，成为许多学者研究的对象。Li 等[8]在 2014 年构建一个新的数据集合，这个数据集包括酒店，餐厅和医院三个不同的领域，且每一个领域的数据都有真实评论、虚假评论以及领域专家发布的虚假评论。利用众包平台雇人撰写的虚假评论，与现实平台中的虚假评论有着较大差异，运用众包平台获取的虚假评论预测现实世界的虚假评论存在一定的偏差。

国外有许多专家通过人工构造数据集，已有相对较为成熟的数据集，但在中文领域并未有相对成熟的人工数据集；而且在语法、构词等方面，中文和英文大不相同，因此不能将训练英文数据集的分类器去预测中文数据集。

在国内文献中，许多学者直接基于英文数据集进行研究，或者利用人工标注的方法获取训练样本。陈燕方等[21]从淘宝网站获取 5000 条中文评论数据，进行专家标注得到一个有标注的数据集，再进行后续研究，但是该数据集数量较少。富越等[19]人通过伪装刷客，潜入第三方刷单平台获取刷客信息，爬取刷客信息在淘宝网站的评论内容，获取虚假评论。王梦华[22]提出利用规则算法与人工标注相结合的方法构建标注训练集，通过机器学习的方法识别重复数据，将重复评论和人工标注集结合起来作为训练集。但是该数据集没有考虑刷单水军团体集中某一时间段内的虚假评论。Lim 等[17]在虚假评论的研究过程中发现虚假评论往往具有突然性、爆发性等周期性的规律，而且这段时间内的评论信息主要是商家为了冲销量进行大规模的刷单产生的评论信息。王禹[23]提出虚假评论

预识别机制，虚假评论结合重复评论识别以及评论时间分布异常情况，构建“虚假评论概率高低”指标进行人工标注。虽然该数据集将重复评论和评论时间异常评论判别为虚假评论，但是人工标注带有主观判断性，也就不可避免的存在误判和漏判的现象。

虚假评论的识别通常作为一个文本分类问题进行建模分析[15]。若使用半监督学习的方法识别虚假评论，训练模型只需要少量标注数据，克服了大量人工标注带来的缺点，同时若在构建数据集时加入机器学习标注的方法，提高标注效率，并一定程度上可以减少主观想法。有些学者仅将一种类型的评论标记为虚假评论，比如重复评论，如此标注样本过于片面，需要从多个方面考虑标注，如集中于某一时间段的大量异常评论、内容较为宽泛、未反应商品特性的评论等。

基于此，本文采用机器学习和专家指导相结合的方法标注数据集：首先利用时间序列的异常检测标注集中于某一时间段（除去电商平台大型活动时间）的潜在评论，其次利用 TF-IDF 方法计算文本评论之间的相似度标注重复评论，再邀请从事多年电子商务领域的专家进行指导，进行数据标注，为后续使用 PU 学习算法识别虚假评论准备样本集。

1.3 研究意义

对于电商平台中海量的评论数据，通过人工识别效率低下，而使用机器学习相关方法快速有效的识别虚假评论，自动过滤虚假评论有着重大的实际应用意义。从海量的电商评论数据中，挖掘与分析虚假评论，提高商品评论的可靠性与真实性，无论是对于电商平台，还是广大消费者，自动过滤虚假评论均拥有深远的意义。有效、低成本、快速地识别电商平台中的虚假评论，对于消费者而言，有助于潜在购买者于购买前期获得更加真实的评论内容，参考评论的内容更加可靠，提高自我判断的能力，并做出正确的购买决策；对于电商平台而言，有助于减少店铺之间的恶性竞争，有助于构建一个公平、真实的购物平台。

1.4 研究内容

本文主要研究现实生活的电商平台中的虚假评论，研究目的是建立一套完善的虚假评论识别体系，从训练数据集的有效构建，到快速准确地识别虚假评论，最后为模型应用分析。

(1) 构建训练数据集，从现实生活获取数据源，利用机器学习和专家标注的方法构建标记数据集，减少人工标注的主观臆断，提高标注效率，为中文虚假评论的标记数据集做出一定的贡献。

(2) 快速准确地识别虚假评论，选择半监督学习领域中的 PU 学习算法，并使用不同的机器学习分类器，从中选择优秀的分类器结合 PU 学习构建分类模型，使用少量的标记样本，自动识别电商平台的文本评论。

(3) 虚假评论应用分析，对电商平台的评论进行预测分类之后，利用词云图展示高频词汇，再通过对比分析主要特征词，体现虚假评论和真实评论之间的内容差异，为消费者提供建议区分虚假评论。并将本文模型应用于其他商品的虚假评论识别，体现模型良好的拓展性。本文的主体框架见图 1.1。

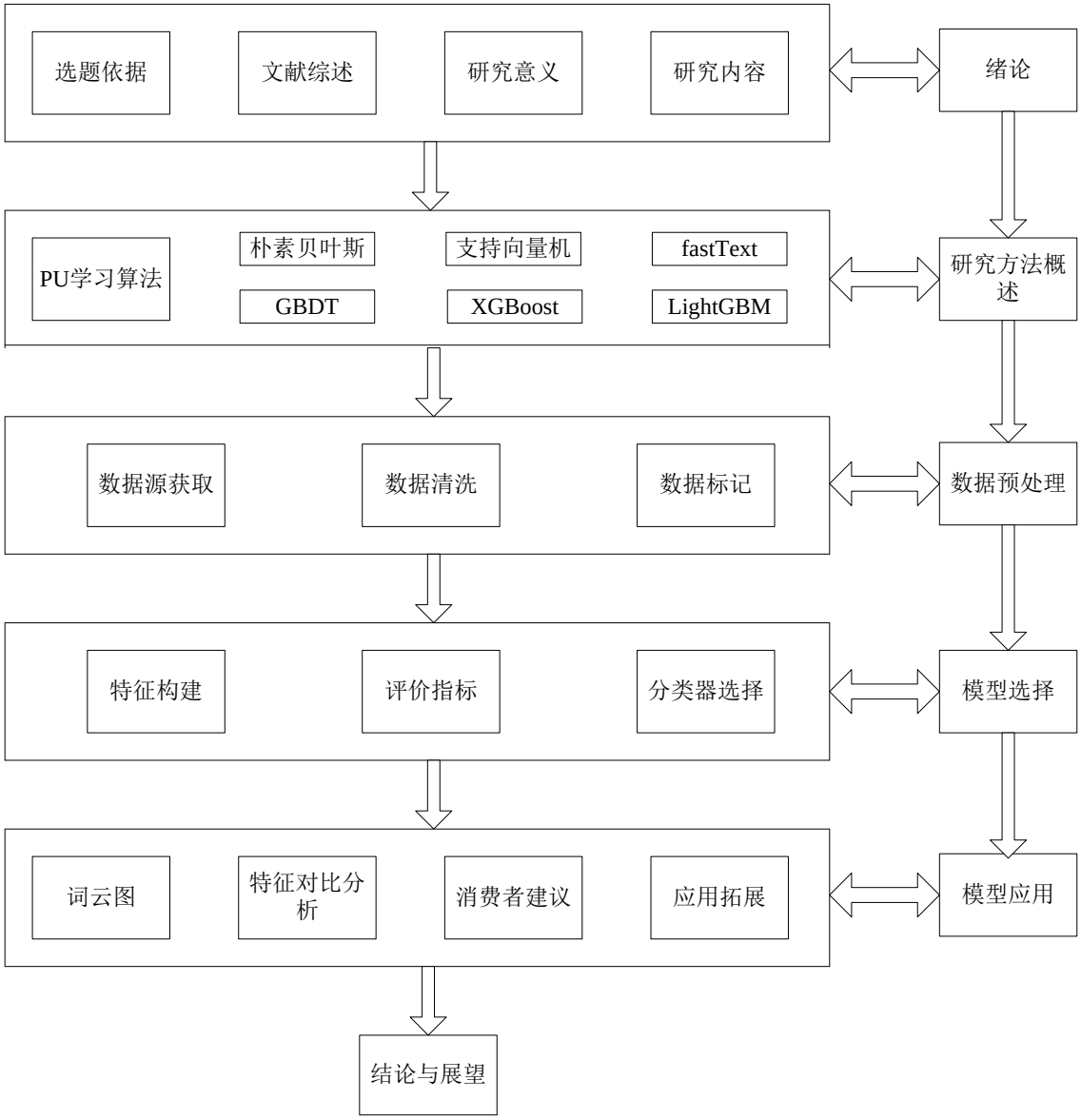

图 1.1 论文主体框架

根据研究思路和主体框架，本文总共分为六部分：

第一章为绪论，主要介绍本文的选题依据，国内外学者对于虚假评论识别领域和数据集构建所做的相关研究并进行相应的评述，再介绍本文的研究意义，主要的研究内容及论文主体框架。

第二章为研究方法概述，基于虚假评论数据集为少量标记数据和大量未标记数据，选用半监督学习领域中的 PU 学习算法的方法进行识别，主要介绍 PU 学习算法的主要思想和应用场景，并介绍了朴素贝叶斯等 6 种分类器的主要思想和优缺点。

第三章为数据预处理，主要介绍本文的数据来源和简单的数据介绍和展示，并介绍

原始文本数据的清洗步骤：删除无关评论、中文分词、去停用词共三个基本步骤，为后续模型应用清洗数据，再详细介绍训练集标注的具体方法。

第四章为模型选择，介绍提取文本特征过程和选择评价指标，运用 PU 学习算法与 6 种分类器进行文本分类，选取效果最优的分类器与 PU 学习相结合，作为最后虚假评论识别的方法。

第五章为虚假评论应用，首先将 PU 学习算法与 LightGBM 应用于虚假评论识别，从整体层面展示虚假评论和真实评论高频词汇的区别，再从具体词汇层面，提取两类评论的 TF-IDF 主要特征词进行对比分析，给消费者提出区分虚假评论和真实评论的建议。其次将模型应用于其他商品的虚假评论识别，能较好的区别虚假评论，体现模型的可拓展性。

第六章为结论与展望，总结本文所做的工作与研究结论，并提出后续可研究方向。

2 研究方法

实际生活的虚假评论识别问题拥有庞大的数据量，使用人工的方法区分虚假评论和真实评论耗时耗力，因此需要使用统计学习的方法快速准确地自动识别虚假评论。若使用监督学习的方法，需使用大量的标记样本，目前缺乏样本集，人工标记效率不高且损耗大量人力物力，因此本文采用半监督学习中较为流行的 PU 学习算法。本章主要介绍虚假评论识别使用的研究方法，首先讲述 PU 学习算法，再简述朴素贝叶斯(naiveBayes)、支持向量机(Support Vector Machines ,SVM)、fastText、GBDT、XGBoost、LightGBM 分类器的基本原理和优缺点，为下一章将 PU 学习算法与 6 种机器学习分类器相结合奠定理论基础。

2.1 PU 学习算法

虚假评论识别问题是分类模型，传统的分类算法大多基于大量已标记数据进行训练学习。基于虚假评论数据集构建的困难性，获取大量标记样本较为艰难，虚假评论分类任务主要从少量标记的正例数据和未标记数据中学习到关键信息构建最终分类模型。本文主要从真实网购平台获取评论数据，使用机器学习和专家标注的方法获取标记数据集，基于少量标记数据和大量未标记数据，本文选取 PU 学习算法来识别虚假评论。

PU 学习算法是一种半监督学习的分类技术，解决了仅有正例和未标记样本的两分类问题[14]。任亚峰等[16]根据对未标注数据集使用情况，将 PU 学习算法分为两类，一是通过正例和未标记数据集中的部分样例来构建分类器，二是使用正例和未标记数据集中的全部样例来构建分类器。与传统常用的监督学习算法不同的是，PU 学习通过少量标记正例样本和大量未标记的样本进行学习，使用少量的标记样本能够达到较好的分类效果。

鉴于 PU 学习算法良好的学习性能，广泛应用于众多方面。针对软件故障数据中正例样本相对较少，而且大量样本标注困难的现实情况，将 PU 学习算法用于软件故障检测研究具有更高的检测率[25]。在实际使用中，生物数据库并非标准数据库，使用 PU 学习算法对位置磷酸激酶抑制剂筛选，输出未标记样本的概率对磷酸激酶抑制剂进行预测，获得较好的预测性能[26]。基于 PU 学习算法，通过信息推送的内容过滤算法，提高了内容过滤的精度和效率[27]。

虚假评论数据集构建困难，需要耗费大量的人力物力财力，而 PU 学习通过少量标

记正例样本和大量未标记的样本进行学习，能取得较好的分类结果，因此在虚假评论领域拥有众多应用[14,28]。Hernández 等[14]使用 PU 学习进行虚假评论的识别，在酒店评论数据集中，对比基准结果(将未标记样例作为负例样本进行模型构建)、单个分类效果(仅使用正例样本构建分类器)与 PU 学习效果发现，PU 学习效果较好。

PU 学习算法的核心思想为：首先从未标记数据中识别一组可信度较高的负例，再基于所有正例和可信负例，使用诸如支持向量机(support vector machine,SVM)等模型算法对未标记的数据进行训练和判别，移除被分类为正例的数据，迭代该过程直到达到某一停止标准。算法流程图如图 2.1 所示。

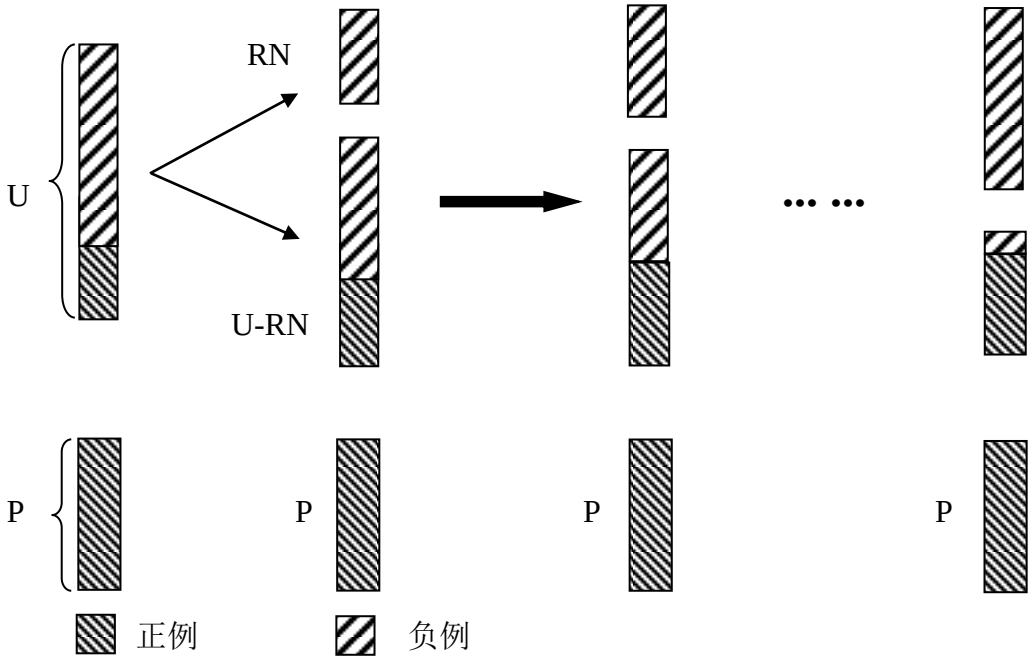

图 2.1 PU **学习算法流程图**

为阐明虚假评论分类器构建的过程，图 2.1 给出 PU 学习算法详细流程图，给定一个训练集 P，其中只含有正例样本，不含有反例样本；未标注数据集 U 中，同时含有正例样本和反例样本，PU 学习算法主要分为两个步骤[24]：

1. 根据已标注的正例样本 P 和未标注数据集 U 进行学习分类，分类器对未标注数据集 U 进行预测，预测为 0 的负例样本定义为可信负例样本集(Reliable Negative Examples，RN)，预测为 1 的负例样本定义为可能的正例样本集，记为 U- RN，将 U 分为 RN 和 U- RN；

2. 利用分类器迭代训练正例样本 P 和可信负例样本集 RN，预测可能的正例样本集

U-RN，直至满足某一特定条件迭代停止，建立最终分类器。

PU 学习算法伪代码如下：

$i \leftarrow 1$

C_i ←Generate Classifier(P, U)

${U_i}^L \leftarrow C_i(U)$

Q_i ←Extract Negatives$({U_i}^L)$

$RN_i \leftarrow Q_i$

$U_i \leftarrow U - Q_i$

while $|Q_i| > 0$ do

$i \leftarrow i + 1$

C_i ←Generate Classifier (P, RN_{i-1})

${U_i}^L \leftarrow C_i(U_{i-1})$

Q_i ←Extract Negatives$({U_i}^L)$

$U_i \leftarrow U_{i-1} - Q_i$

$RN_i \leftarrow RN_{i-1} + Q_i$

Return (C_i)

2.2 分类器

在上述小节中主要讲述 PU 学习算法的原理及流程。在本节将介绍朴素贝叶斯、支持向量机、fastText、GBDT、XGBoost、LightGBM 共 6 种分类器的算法主要思想及对应的优缺点，为便于与 PU 学习算法进行结合。

2.2.1 朴素贝叶斯

朴素贝叶斯分类思想是假定对给出的待分类项，通过求解在此项出现的条件下各个类别出现的概率，选取概率最大的待分类项判定为该类别。

朴素贝叶斯计算流程如下：

(1) 设$x = \{a_1, a_2, \ldots, a_m\}$为一个待分类项，$a$表示$x$的一个特征属性。

(2) 类别集合$C = \{y_1, y_2, \ldots, y_n\}$。

(3) 计算$P(y_1|x), P(y_2|x), \ldots, P(y_n|x)$。

其中$P(y_1|x), P(y_2|x), \ldots, P(y_n|x)$可利用如下步骤进行计算：

(1) 选出一个已知分类的待分类项集合，把该集合作为训练集。

(2) 计算在每个类别下各个特征属性的条件概率估计。即

$$\begin{aligned} &P(a_1|y_1), P(a_2|y_1), \ldots, P(a_m|y_1) \\ &P(a_1|y_2), P(a_2|y_2), \ldots, P(a_m|y_2) \\ &\ldots \\ &P(a_1|y_n), P(a_2|y_n), \ldots, P(a_m|y_n) \end{aligned} \tag{2.1}$$

(3) 若各个特征属性是条件独立的，则根据贝叶斯定理有以下推导：

$$P(y_i|x) = \frac{P(x|y_i)\,P(y_i)}{P(x)}$$

其中

$$\begin{aligned} &P(x|y_i)\,P(y_i) \\ &= P(a_1|y_i)\,P(a_2|y_i)\ldots P(a_m|y_i)\,P(y_i) \\ &= P(y_i)\prod_{j=1}^{m} P(a_j|y_i) \end{aligned} \tag{2.2}$$

(4) 如果$P(y_k|x) = \max\{P(y_1|x), P(y_2|x), \ldots, P(y_n|x)\}$则$x \in y_k$。

朴素贝叶斯的主要优点有：

(1) 朴素贝叶斯模型分类效果比较稳定；

(2) 小规模数据分类效果较好，也能够处理多分类任务；

(3) 对缺失数据不太敏感，算法较为简单。

朴素贝叶斯的主要缺点有：

(1) 理论上朴素贝叶斯模型与其他分类方法相比误差率较低。但是实际上并非如此，因为理论上假设属性之间相互独立，但这个假设在实际应用中往往是不成立的；

(2) 需要预先知道先验概率，会由于选错先验模型导致预测效果不好；

(3) 受先验概率和数据的影响得到的后验概率作为分类标准会存在一定的错误率；

(4) 对数据形式很敏感。

2.2.2 支持向量机

支持向量机是一种监督学习算法，在分类以及回归分析中有着广泛地应用。支持向量机属于一般化线性分类器，该类分类器能够同时将经验误差最小化与将几何边缘区最大化，因此也被称为最大边缘区分类器。其核心思想是寻找一个涵盖线性可分支

持向量机、线性支持向量机、非线性不可分向量机三大类问题。下面主要就支持向量机的优缺点进行阐述。

支持向量机的优点：

(1) SVM 利用内积核函数代替向高维空间的非线性映射，可灵活使用各种核函数来处理分类和回归问题；

(2) 支持向量在 SVM 分类决策中起决定作用，少数支持向量确定了最终的决策函数，无需使用全部数据，有较好的“鲁棒”性。

支持向量机的缺点：

(1) 由于 SVM 是借助二次规划来求解支持向量，而二次规划涉及m阶矩阵的计算，当m很大时矩阵存储和计算将耗费大量的计算机内存和运算时间；

(2) 传统的支持向量机算法只给出了二类分类的算法，但在实际应用中，需要解决多分类问题。

2.2.3 fastText 算法

fastText 是一种 Facebook AI Research 在 16 年开源的一个文本分类器。fastText 的核心思想是将整篇文档的词及 n-gram 向量叠加平均得到文档向量，然后使用文档向量做 softmax 多分类。其中 n-gram 是一种基于语言模型的算法，基本思想是将文本内容按照字节顺序进行大小为N的分割，最终形成长度为N的字节片段序列。softmax 是逻辑回归在多分类问题上的一个推广。其损失函数形式如下

$$J(\theta) = -\left[\sum_{i=1}^{m} y^{(i)} \log h_\theta\left(x^{(i)}\right) + \left(1 - y^{(i)}\right) \log\left(1 - h_\theta\left(x^{(i)}\right)\right)\right] \tag{2.3}$$

其中$\left\{\left(x^{(1)}, y^{(1)}\right), \ldots, \left(x^{(m)}, y^{(m)}\right)\right\}$表示$m$个被标注的样本，$x^{(i)} \in R^n$，且$y^{(i)} \in \{0, 1\}$。链接函数$h$与上节相同。

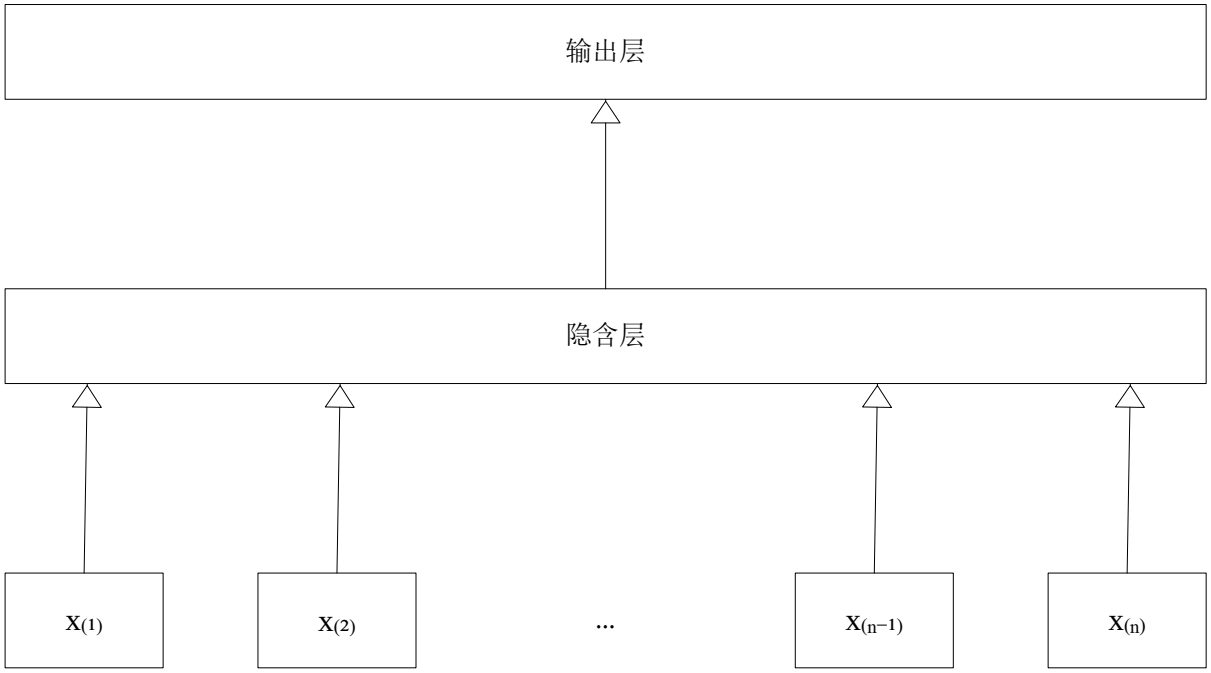

图 2.2 fastText **算法模型架构图**

fastText 算法优点：

(1) 适合大型数据+高效的训练速度：能够训练模型“在使用标准多核 CPU 的情况下 10 分钟内处理超过 10 亿个词汇”；

(2) 支持多语言表达：利用其语言形态结构，fastText 能够被设计用来支持包括英语、德语、西班牙语、法语以及捷克语等多种语言。fastText 的性能要优于 word2vec 工具，也优于其他目前先进的词态词汇表征；

(3) fastText 专注于文本分类，在许多标准问题上实现当下最好的表现(例如文本倾向性分析或标签预测)。

2.2.4 GBDT 算法

迭代决策树(Gradient Boosting Decison Tree，GBDT)属于集成学习。集成学习[29]通过使用多个基学习器(分类或回归模型)进行学习，按照某特定规则进行组合，以期获得效果更好的模型。目前主要的集成学习有Bagging (bootstrap aggregating)和Boosting，Bagging[30]是一种并行的集成学习，同时进行基学习器的学习，其主要思想是利用自助抽样法(bootstrap)[31]对训练集进行抽样产生新的训练集，在每个新训练集上构建一个学习器，再对各个学习器进行组合(投票或平均)得到最终的预测模型。Boosting [32]是一种

串行的集成学习，基学习器之间存在依赖关系，每个学习器依赖之前学习器的训练结果，集中关注被错分的数据以获取新的学习器，赋予学习器不同的权重进行组合得到最终的预测模型。

GBDT 是 2001 年 Jerome Friedman 提出的一个 Boosting 算法[33]，是一种迭代的决策树算法，每颗决策树训练在之前决策树训练结果中的错误，以第二棵决策树训练的目标为例，第二棵树的训练目标为第一棵决策树的训练结果与真实值之间的残差，加和每一棵决策树的结果得到最终结果。

GBDT 算法的具体流程如下：N为样本数，M为迭代次数

(1) 初始化弱学习器

$$f_0(x) = \arg\min \sum_{i=1}^{N} L\left(y_i, c\right)$$

(2) 对迭代次数$m = 1, 2, ..., M$有：

a) 对每个样本$i = 1, 2, ..., N$，计算负梯度，即残差

$$r_{im} = -\left[\frac{\partial L\left(y_i, f\left(x_i\right)\right)}{\partial f\left(x_i\right)}\right]_{f(x)=f_{x-1}(x)}$$

b) 将残差作为样本新的真实值，且将数据(x_i, r_{im})，$i = 1, 2, ..., N$作为下棵树训练数据，得到新的回归树$f_m(x)$，其对应的叶子节点区域为R_{jm}，$j = 1, 2, .., J$，其中J表示回归树t的叶子节点个数。

c) 对叶子区域$j = 1, 2, .., J$，计算最佳拟合值

$$\gamma_{jm} = \arg\min_{\gamma} \sum_{x_i \in R_{jm}} L\left(y_i, f_{m-1}\left(x_i\right) + \gamma\right)$$

d) 更新强学习器

$$f_m(x) = f_{m-1}(x) + \sum_{j=1}^{J} \gamma_{jm} I\left(x \in R_{jm}\right)$$

(3) 最终学习器

$$f(x) = f_M(x) = f_0(x) + \sum_{m=1}^{M} \sum_{j=1}^{J} \gamma_{jm} I\left(x \in R_{jm}\right)$$

GBDT 的优点：

(1)预测精度较高;

(2)能灵活处理各种类型的数据。

缺点：

(1)弱学习器之间存在依赖关系，难以并行；

(2)当数据维度较高时，计算复杂度会加大。

2.2.5 XGBoost 算法

XGBoost[34]是在 Gradient Boosting 决策树基础上发展起来的，传统的 GBDT 只利用一阶的导数信息，而 XGBoost 对损失函数进行二阶的泰勒展开，求得到的最优解的效率更高[35]，且加入正则化函数。XGBoost 算法将决策树与 GBDT 算法进行结合，三者关系如下图所示：

图 2.3 XGBoost **算法模型架构图**

XGBoost 核心思想是不断地添加二叉树，多次通过特征分裂来生长一棵树，每次添加一个树，通过稀疏感知算法进行学习。其学习过程是建立一个新函数去拟合上一步预测的残差。假设训练完成后得到k棵树，需要根据样本叶子节点的特征预测样本的分数，最后将每棵树下得到的分数相加得到该样本的预测值。

XGBoost 主要优点如下：

(1)精度高，XGBoost 损失函数用到了二阶导数信息，这样就为损失函数指定了梯度方向，从而提高模型精度和速度；

(2)速度快，可以利用 CPU 进行多线程运算；

(3)可扩展性强，其目标函数支持 linear、logistic、softmax 等，可以处理回归、二分类，多分类问题。另外还可以自定义损失函数；

(4)在损失函数里加入正则项，防止过拟合。

缺点：

(1) XGBoost 的精确贪心算法在每轮迭代时，都需要遍历整个训练数据多次。反复地读写训练数据又会消耗非常大的时间;

(2)Level-wise 迭代方式可能产生不必要的叶结点;

(3)适用于高偏差，低方差的训练集。

2.2.6 LightGBM 算法

LightGBM (Light Gradient Boosting Machine)算法由 Ke，Meng 等人[36]提出，基于 GBDT 算法的框架进行优化。LightGBM 使用 Leafl-wise 按叶节点分裂的方法，在决策树分裂的过程中，遍历所有的节点，选择最大增益的叶节点分裂，加快算法的运行速度，按叶分裂的决策树过程如图 2.4 所示。

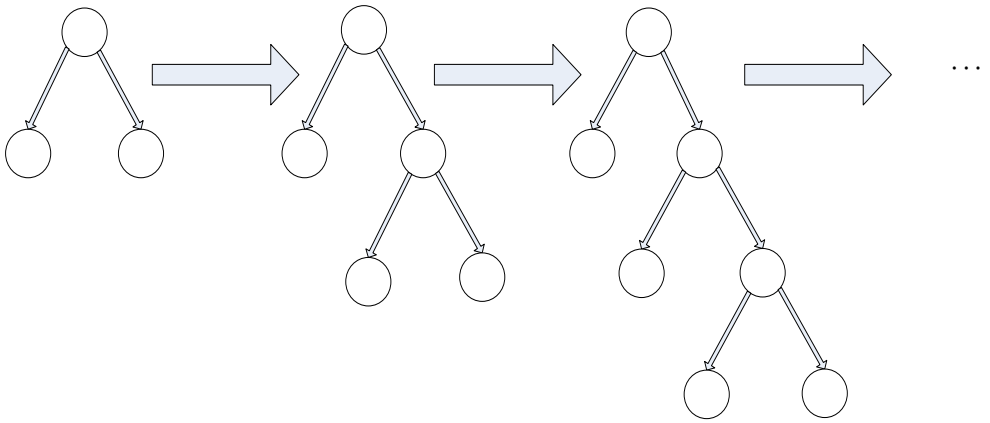

图 2.4 按叶分裂的决策树过程

LightGBM 算法内部使用直方图(histogram)算法，其思想是将连续的浮点特征离散成k个离散值，并构造宽度为k的 histogram，通过遍历训练数据，统计每个离散值在直方图中的累计统计量。在进行特征选择时，只需要根据直方图的离散值，遍历寻找最优的分割点。

LightGBM 优点：

(1)基于直方图的稀疏特征优化；

(2)多线程优化；

(3)更快的训练效率；

(4)低内存使用；

(5)更高的准确率；

(6)支持并行化学习；

(7)可处理大规模数据；

(8)支持直接使用分类特征。

2.3 本章小结

本章节主要介绍了 6 个分类器，首先介绍了两个传统机器学习算法朴素贝叶斯、支持向量机，用于文本分类的 fastText，及目前先进的集成算法 GBDT、XGBoost、LightGBM，并对这些分类器进行简要介绍。在这一章主要介绍各种分类器的原理及对应的优缺点，在后续模型构造部分将这些分类器与 PU 学习算法进行结合。

3 数据预处理

本文主要研究实际生活中真实的虚假评论问题，本章描述数据获取的来源和数据预处理的过程，并介绍虚假评论标记数据集的构建，为后续模型识别虚假评论做好前期数据准备。

3.1 数据源获取

本文评论数据来源国内某大型电商平台，利用 python 网络爬虫技术爬取某品牌运动鞋中一款热销鞋子的评论，评论时间跨度从 2017 年 6 月 5 日至 2019 年 12 月 13 日，共采集到 21133 条原始数据，包括商品名称、颜色、尺码、评论时间、评论内容、评分、图片数量、追评数量、回复追评数量共 9 个维度数据，表 3.1 为数据样例：

表 3.1 评论数据样例

商品名称	颜色	尺码	评论时间	评论内容	评分	图片数量	追评数量	回复追评数量
XX 男鞋跑步鞋舒适透气运动鞋	黑色/银色	41	2017/6/5	穿着很轻，很舒服，质量一般	5	0	0	0
XX 男鞋跑步鞋舒适透气运动鞋	黑色/银色	42	2017/6/5	穿上试了一下，感觉不错，挺舒服，很轻，外观也很漂亮，第三双乔丹了，之前两双质量都非常好穿两年多一点都没坏，希望这个质量一样好	5	2	0	0
XX 男鞋跑步鞋舒适透气运动鞋	深藏青/闪亮橘	41	2017/6/7	穿了两天，比较舒适。质量还可以。	5	2	0	0
XX 男鞋跑步鞋舒适透气运动鞋	黑色/银色	44.5	2017/6/9	不错，非常满意的一次购物，给老公买的，老公非常喜欢，说穿上舒服轻巧，透气型也强，一次买了两双，穿穿得劲会再买的。	5	4	1	0
XX 男鞋跑步鞋舒适透气运动鞋	黑色/银色	45	2017/6/9	很好不错，快递很快。昨天买的今天就到了。	5	0	0	0

本文选择的这款热销运动鞋，在该电商平台上共销售 7 个颜色，将采集到的原始评论按照颜色进行分类汇总，得到表 3.2。“黑色/银色”的运动鞋拥有最多的评论数量，占总评论数的约 50%，而“学院蓝/闪亮橘”的运动鞋评论有约 1.58%。

表 3.2 评论数量按商品颜色分布

颜色	汇总	占比
黑色/极光红	2873	13.595%
黑色/极光红（革面）	838	3.965%
黑色/银色	10581	50.069%
黑色/银色（革面）	1434	6.786%
深藏青/闪亮橘	4490	21.246%
学院蓝/闪亮橘	333	1.576%
学院蓝/闪亮橘（革面）	584	2.763%
总计	21133	100%

电商平台上对于商品的评分共分为 5 档，5 分是最高分，1 分表示最差，图 3.1 根据原始评论的评分进行汇总统计，发现 96%的购买者给了 5 分好评。

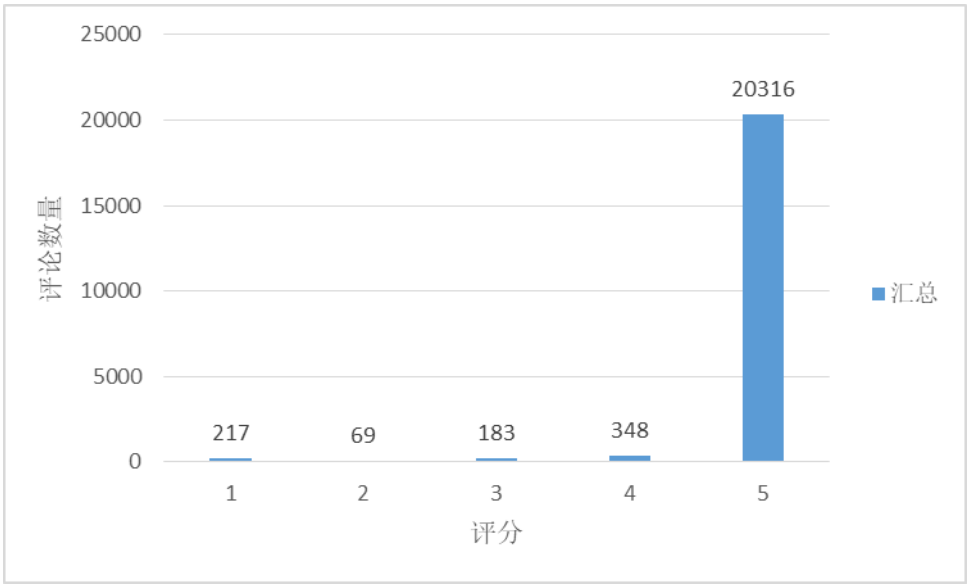

图 3.1 评论数量按评论分布

评论数据的时间跨度为 2017 年 6 月到 2019 年 12 月，图 3.2 从评论日期的角度进行评论汇总。

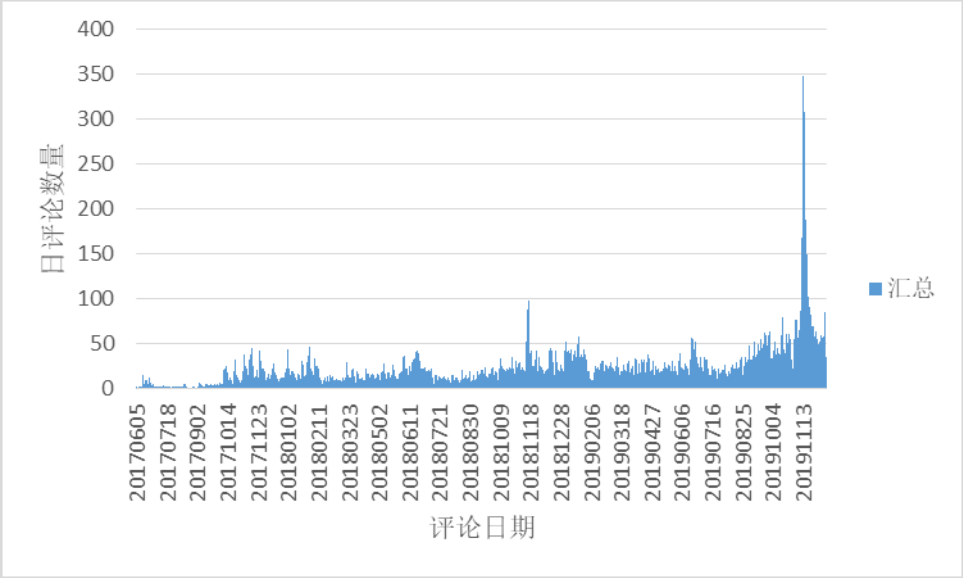

图 3.2 评论数量按评论日期分布

图 3.2 中可以发现在每年的“双十一”(11 月 11 日)前后几天评论数据有一定突增，尤其是在 2019 年“双十一”评论数量远高于前两年，可能是今年双十一活动优惠力度较大或前期宣传效果较好，吸引众多顾客前来购买并进行商品评价。

3.2 文本数据清洗

原始文本评论来源于电商平台已发布的评论，获取的评论不能直接用于模型构建，需要进一步对文本数据进行预处理，主要通过剔除特殊字母、数字、标点符号以及诸如“系统默认好评”、语句内部无语言逻辑等无关评论。预处理完成后进行分词，分词部分采用 python 内置的 jieba 分词模块来实现中文分词，并剔除文本词语中一些无实际意义的词语或符号。

3.2.1 无关评论删除

在采集到的原始商品评论中，经统计存在 483 条默认好评，如“您没有填写内容，默认好评”、“此用户未及时填写评价内容”、“此用户未及时填写评价内容，系统默认评价！”、“此用户未及时评价，系统默认好评。”，默认好评无实际含义和研究意义，在

后续研究中给予删除。

本文的数据来源于真实顾客的评论，有些顾客发表的评论为无关评论，无关评论中包含的信息内容少且不易进行特征提取。为后续研究创造相对干净有效的数据，需要剔除无关评论，剔除评论文本中仅有标点符号或者数字、全为标点符号和数字的组合、语句内部无语言逻辑、不重复内容长度小于 3 的评论，部分评论示例及评论数量如表 3.3 所示。经统计无关评论共有 809 条，经过初步数据清洗之后共有 19840 条数据。

表 3.3 无关评论

无关评论类型	评论示例	评论数量
仅有标点符号或者数字	“，。？！！？？？？？？？”	72
全为标点符号和数字的组合	“σσqnreadingpast”	17
语句内部无语言逻辑	“KTV 哦哦 YY 五 KKK 啊 YY9 哦图”	44
不重复内容长度小于 3	“哈哈哈哈哈哈哈哈哈哈哈哈哈哈哈”	676

3.2.2 中文分词

与英文不同，中文两个词之间没有空格，因此在利用计算机处理中文时，面临的首要问题就是对中文进行分词问题。目前中文自动分词面临的主要难题有：分词规范、歧义切分和识别未登录词。当前使用的中文分词方法大致分为以下三种：基于字符串匹配的机械分词方法、基于理解的分词方法和基于统计词频的分词方法。本文采用 python 内置的 jieba 分词模块来实现中文分词[37]。

以随机获取一条原始评论为例，展示其分词结果。原始评论为：“虽然不是很喜欢这个款式，但鞋弹性非常好，穿着很舒服！很透气！”经过 jieba 分词得到的结果为：['虽然'，'不是'，，'很'，'喜欢'，'这个'，'款式'，'，'，'但'，'鞋'，'弹性'，'非常'，'好'，'，'，'穿着'，'很'，'舒服'，'！'，'很'，'透气'，'！']。由jieba分词模块的分词结果发现，分词是符合中文词语结构和语言情景的，为后续将文本转换为文本向量打下了良好的基础。

3.2.3 去停用词

经过中文分词之后，发现有许多无实际含义的词语，如 “的”、 “等”、 “且”、 “还有”的连词、介词等词语，还有标点符号和英文字符，把这些词语称为“停用词”。在常

见的中文文本预处理的步骤中，需要剔除停用词。去停用词可以降低数据维度，并且能快速获取到关键信息，本文选择哈工大停用词表，并根据本文数据的现实情况添加额外的停用词，使用的停用词示例见表 3.4。

表 3.4 停用词示例

bull	forall	hellip	nabla	nagaitain	lowast
▼	_&	〖	〗	。。	\
哎呀	啊	的	不然	尽管	还有

3.3 训练集构建

目前众多学者在虚假评论或者垃圾评论识别领域多采用国外具有代表性的公开数据集，在中文评论领域缺乏成熟完整的标记数据集，因此需要构建一个真实的中文训练数据集。

目前众学者研究过程中使用的数据集主要分为两大类：众包平台评论和商业评论数据[20]。第一类是众包平台评论，来自各大众包平台，雇佣人员发表虚假评论，得到已经标注完成的数据集，但是该方法花费成本较高，且与现实评论相距甚远。第二类是商业评论数据，来源于现实生活中的购物平台或点评网站，但是真实评论与虚假评论相互掺杂，大量标记数据集的成本较高。目前研究者们主要使用“黄金标准数据集”、Yelp 评论、TripAdvisor 评论等英文数据集，缺乏成熟的虚假评论中文数据集。

本文为挖掘与分析现实生活的虚假评论，现已获取电商平台中消费者产生的评论，需要构造标记数据集，为后续模型构建做准备。通过咨询某电商公司拥有丰富工作经验的专家，根据专家反馈，一般虚假评论发表的时间较为集中；评论内容有着较高的相似性；评论内容的情绪较为高涨，存在一定的鼓动性等。因此本文由三个电商专家从数据集中标记虚假评论，再通过机器学习的方法，构建虚假评论候选数据集，选取时间异常的重复评论标记为虚假评论。结合机器学习和专家标注的方法进行数据集标注，可以提高一定的标注效率和准确性。

3.3.1 重复评论

重复评论是指不同用户在购买同一款商品后，发表有高度相似度的评论。本文通过

文本评论相似度的计算来识别重复评论，若两句评论之间的相似度大于等于 0.9，则判定为重复评论，并将其归入标记数据集的候选评论。

本文使用由 Radim Rehurek 编写的 gensim 库③，gensim 库主要用于计算文本相似度，涉及语料库的建立、词频-逆文本频率(Term Frequency–Inverse Document Frequency，TF-IDF)模型[38, 39]和计算余弦相似度[40]。

(1)词频-逆文本频率(TF-IDF)模型

TF-IDF 是一种统计方法，主要用于计算文本特征的权重。该方法结合词频和逆文本频率，从文档中出现词语的频率以及词语总体分布两个方面入手，去衡量词语的重要性。如果一个词语在该文档中重复出现，而在其他文档中较低频率的出现，则说明该词语与这个文档关联性更紧密，能够较好的区分类别。TF-IDF 方法的计算公式如下：

$$w_{ij} = tf \times idf = \frac{n}{N_i} \cdot \log_2 \frac{N}{m} \tag{3.1}$$

其中，w_{ij}表示文本d_i中第j个词的权重。N表示文本数据集总数量，m表示含有词语t_{ij}的文本数量，N_i表示文本d_i中所有词语出现的总数，n表示指词语t_{ij}在文本d_i中出现的次数。词频tf表示某个词语在一个文本中出现的频数，频数越大，表示该词语对文本的贡献度越大。逆文本频率idf表示词语在所有文本数据集中的分布情况，包含该词语的文本数目越少，idf则越大，说明这个词语越适合分类。

(2)余弦相似度

余弦相似度是计算空间中两个向量之间的夹角，来判断两个向量之间的相似程度。若两个向量之间的夹角越大，则向量代表的文本相似度越小；若两个向量之间的夹角越小，则向量代表的文本相似度越大。

假设文本A和文本B是两个n维向量，则两个向量之间的余弦值为：

$$\cos\theta = \frac{\sum_{i=1}^{n}(\mathrm{A}_i \times \mathrm{B}_i)}{\sqrt{\sum_{i=1}^{n}(\mathrm{A}_i)^2} \times \sqrt{\sum_{i=1}^{n}(\mathrm{B}_i)^2}} \tag{3.2}$$

当余弦值越接近于 1，说明向量之间的夹角接近 0 度，两个文本内容就越相似。

根据上述理论知识的介绍，计算评论相似度的过程如下：

(1)去除无效和无关评论、中文分词、剔除停用词；

(2)构建评论的语料库，将每条评论转换为向量；

(3)训练评论的 TF-IDF 模型；

③ https://radimrehurek.com/gensim/

(4)利用余弦相似度，计算文本与文本之间相似度。

根据评论间的相似度，取出相似度大于等于 0.9 的评论，挑选出重复评论，作为虚假评论数据集的候选数据集。

3.3.2 时间序列异常评论

根据电子商务领域的专家反馈，电商评论的商品销售量和评论数量，除去大型促销活动节日(如“618”、“双十一”、“双十二”)，总体趋势是保持平稳的。有学者研究发现虚假评论有一定的时间集中性，一般多在一定时期内涌现[41]，因此基于评论发表的时间分布来提取标记数据集的候选评论。

图 3.2 的评论时间横跨 2017 年至 2019 年，观察总体评论数量趋势变化，而发现 2017 年和 2019 年的总体评论数量差异较大，使 2017 年评论分布较不明显，因此按年拆分分别进行汇总，如图 3.3，图 3.4，图 3.5 所示。

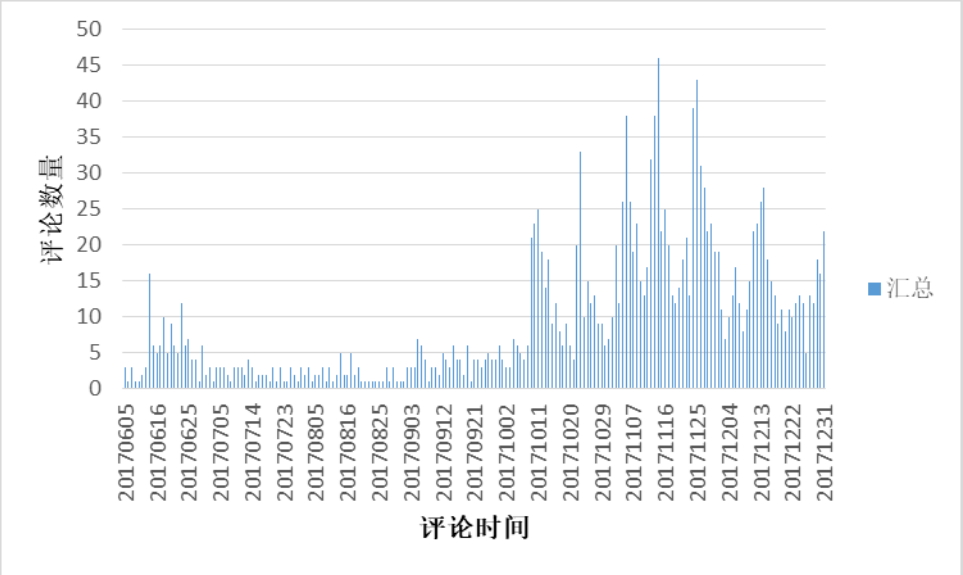

图 3.3 2017 年评论数量按日分布

从图 3.3 可以看出，日评论数量在 2017 年 6 月中旬有小幅度增加，但在 10 月、11 月、12 月其中几天的日评论数量有大幅度提升，期间有“双十一”和“双十二”两个促销时间段会使得销量增加且活动后的日评论数据增多，但是非促销时间段的日评论数量增加是异常现象，可能为商家雇佣刷单机构进行集中刷单。

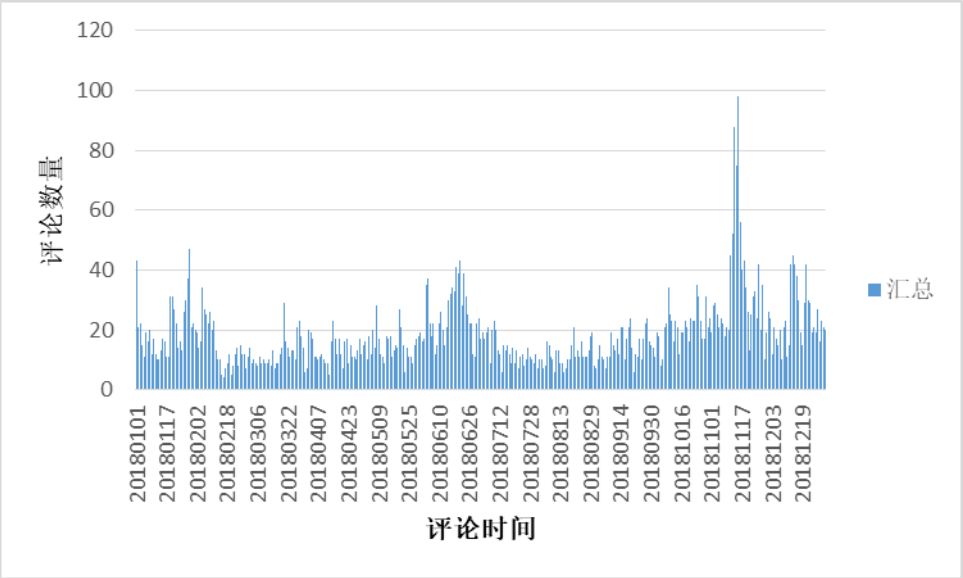

图 3.4 2018 年评论数量按日分布

观察图 3.4 发现，除去特殊促销活动期间，在 2018 年 1 月初、2 月初、6 月中旬日评论数量有相对的增加，在后续数据标注的时候重点关注这些异常时间段内的评论。

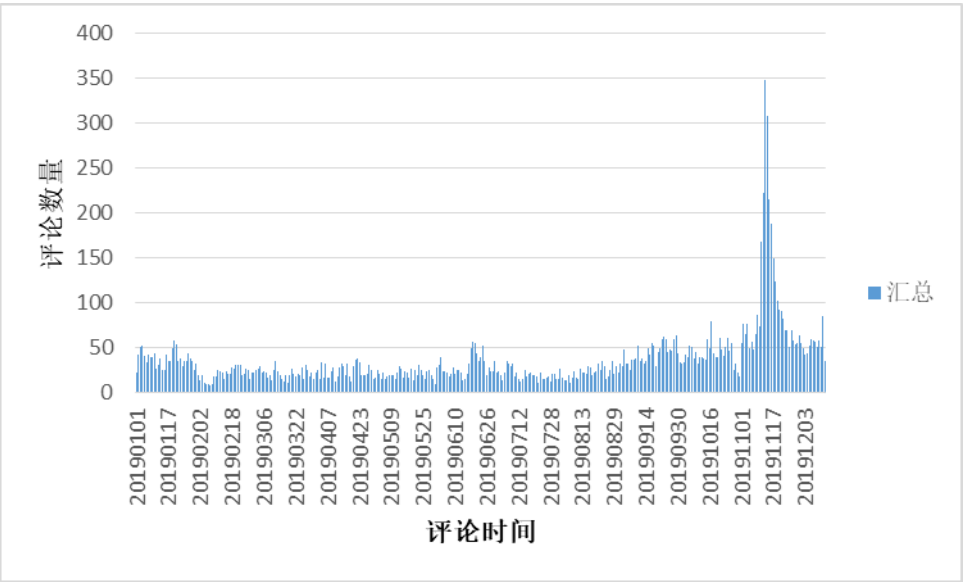

图 3.5 2019 年评论数量按日分布

图 3.5 显示，2019 年日评论数量总体比较稳定，仅在 11 月中旬有大幅度的增加，从评论时间分布上看，2019 年评论异常时间段仅为“双十一”活动期间。

从图 3.3，图 3.4，图 3.5 观察到评论数量在日分布的总体变化趋势，还需对日评论数据进行描述统计，如表 3.5 所示。从日趋势图中不能从具体数值上发现当日评论数异

常，因此需定义日评论数的异常值[④]，本文将与平均值的偏差超过两倍标准差的值定义为异常值。

表 3.5 日评论数量的描述统计

指标	2017 年	2018 年	2019 年	总体
平均值	9.186	18.293	36.386	23.198
1/4 分位数	3	11	20	11
3/4 分位数	13	22	42	28
最大值	46	47	349	349
最小值	1	4	8	1
标准差	9.147	10.923	33.890	24.948
异常值阈值	27.479	40.139	104.167	73.093

由于 2017 年至 2019 年三年的日均评论数量差异较多，因此按年分别进行挑选。经统计，2017 年评论超过异常值共有 356 条评论，2018 年评论超过异常值共有 844 条评论，而 2019 年异常评论均于“双十一”期间故不考虑，因此将 2017 年和 2018 年的评论时间异常的评论作为虚假评论集的候选数据集，为后续数据标注做准备。

3.3.3 数据标注

文本为排除人工标注的主观判断性，采用机器学习判定和专家指导相结合的方法进行虚假评论的标注工作。邀请三位从事多年电商工作的专家，结合多年电商工作中对于虚假评论的理解，从虚假评论高重复、时间点异常、情绪异常、内容与商品不符等多方面特性进行数据标注，选出虚假评论（标记为 1）和真实评论（标记为 0）。为克服单人标注的主观判断，遵循少数服从多数的原则，确定评论数据的最终标记类别。

如下展示两条专家标记为虚假评论的样例，将电商平台和鞋子品牌名称用 XX 代替：

(1) 商品不错，性价比很高，外观漂亮大气，质量看起来不错，值得推荐大家购买，一次愉快的购物体验！

(2) 和专卖店的买的一模一样，质量很好。比**上买的同价格的鞋强多了。信赖 XX，支持 XX，物流超快。

除专家标注数据外，再从候选数据集中选出时间异常的重复评论。结合 3.3.1 节发

[④]郑家亨．统计大辞典：中国统计出版社，1995 年 03 月

表时间异常评论的识别发现，有些评论发表的时间间隔时间很短，且评论内容的相似度极高，由 3.3.2 节判断为重复评论，时间异常的重复评论样例展示如表 3.6：

表 3.6 评论时间极度集中且相似度极高的评论实例

发布评论日期	评论内容
2017/10/22 22:36	鞋子很轻 上脚舒适 尺码标准 质量不错
2017/10/22 22:38	鞋子很轻 上脚舒适 尺码标准 质量不错 价格满意
2018/05/20 20:12	鞋子非常合脚，穿着挺舒适的，一下子买了两双一样的，感觉特别酷，比起实体店便宜，质量信得过，放心！！！
2018/05/20 20:12	鞋子非常合脚，穿着挺舒适的，一下子买了两双一样的，感觉特别酷，比起实体店便宜，质量信得过，放心！！！
2018/11/19 23:05	鞋子很好质量不错物有所值，下次还会来的
2018/11/19 23:06	鞋子很好质量不错物有所值，下次光临本店
2019/01/06 21:33	薄款透气耐脏，活动一次买了三双。质量没问题，价格实惠多了。
2019/01/06 21:34	薄款透气耐脏，活动一次买了三双。质量没问题，活动前一个小时抢的。价格实惠多了。

电商专家指出商家刷单行为往往在短时间内完成，多数的虚假评论多集中于某一定时间段，要在短时间内发表大量评论，评论内容会有很高的相似度，而真实评论有着不同的文风和内容，因此在较短时间内发布的重复评论标记为虚假评论。

将专家标注结果和时间异常的重复评论，最终标记为虚假评论，共标记 834 条，为下文 PU 学习算法模型选择与应用奠定数据基础。

3.4 本章小结

本文主要研究实际生活电商平台中的虚假评论，故以商品真实评论出发展开研究。本章具体介绍了数据来源、数据清洗步骤及数据标记过程。首先交代数据的来源，并对数据做简单介绍分析，其次将爬取的数据进行清洗，删除无关评论，中文分词及去停用词；最近介绍数据集标注，利用机器学习与专家标注共同完成虚假评论数据的标记工作，为后续 PU 学习模型的构建完成数据准备工作。

4 模型选择

在第 3 章节简要的对 6 种分类器进行介绍并分析了对应的优缺点，本章节欲将PU学习算法与第 3 章节提到的分类器进行结合，应用于虚假评论的识别模型。Narayan 等[42]已完成决策树、朴素贝叶斯、支持向量机、KNN、随机森林、逻辑回归分类器与 PU 学习算法的结合识别虚假评论，而对于目前较前沿的一些分类器如 fastText、LightGBM、XGBoost 算法等还未涉及。基于此，本文不仅将PU算法结合朴素贝叶斯、SVM传统机器学习算法，还结合目前较流行的 fastText、GBDT、LightGBM、XGBoost 算法，选择最优分类器与 PU 学习算法组合学习进行虚假评论的识别。

4.1 特征构建

文本分类问题是自然语言处理领域中非常经典的问题之一。文本分类问题方面的研究最早可以追溯利用专家规则(pattern)进行分类，但是专家规则分类存在一定弊端，例如：费时费力、覆盖的范围以及准确率有限等缺点。后来随着统计学习方法的发展，尤其是互联网技术的发展，促使网络文本数量呈指数级的增长，与此同时机器学习领域也快速兴起。这些发展使得研究者将机器学习方法广泛应用于文本分类。

常用的机器学习分类方法通过将整个文本分类问题拆分为特征工程和分类器两部分，本文使用的 PU 学习算法和应用的分类器已在第三章进行介绍。特征工程又可细分为文本预处理、特征提取、文本表示三个部分，这种细分的最终目的是将文本转换成计算机可理解的形式，并进一步对特征工程进行封装，为下一步进行分类提供数据支撑，文本预处理和文本表示已在第二章完成，本节主要介绍特征提取。

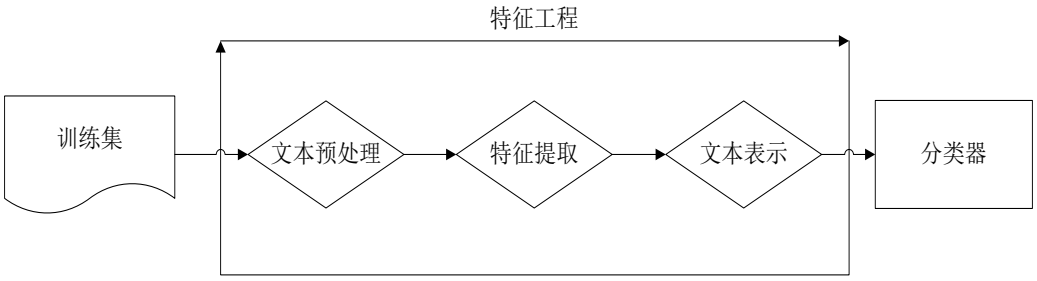

图 4.1 传统机器学习分类方法流程图

本文数据源获取过程中，采集了商品名称、颜色、尺码、评论时间、评论内容、评分、图片数量、追评数量、回复追评数量共 9 个维度数据，欲将评论时间、评论内容、评分、图片数量、追评数量、回复追评数量共 6 个维度进入模型。观察数据发现：图片数量、追评数量、回复追评数量这三个维度数据存在许多 0，对照商品评论原网页发现并不是原始数据存在缺失，而是真实顾客进行评论时并没有发表图片和追加评论，多数顾客仅在购买后发表文字评论。本文主要研究评论内容来识别虚假评论，但是评论内容为短中文文本，不能直接进入模型学习，需要进行特征处理，在第二章数据预处理中，已经介绍中文分词、去停用词、词向量转换，本小节着重介绍提取评论文本的特征，使用的特征汇总如表 4.1 所示：

表 4.1 特征名称

特征序号	特征名称
1	TF-IDF 值
2	评论长度
3	是否包含煽动性词语
4	是否提及电商平台名称
5	是否提及品牌名称
6	评论包含图片数量
7	追评数量
8	回复追评数量

(1) TF-IDF 值：TF-IDF 是一种常见的文本特征提取方法，结合词频和逆文本频率，提取每句评论中的重要词汇，若词语的 TF-IDF 值越大，则说明该词对整体评论越重要，区分度越好。以原始评论“鞋子很好总体上穿着也还算舒服，物流足够快，唯一的遗憾是，左脚脚后跟有一点点的磨脚”为例，经过分词和删除停用词后得到：['鞋子'，'很'，'好'，'总体'，'上'，'穿着'，'还'，'算'，'舒服'，'物流'，，'足够'，'快'，'唯一'，'遗憾'，'左脚'，'脚后跟'，'一点点'，'磨脚']，计算每个词语对应的 TF-IDF 值，具体数值如表 4.2 所示，求得每句评论中每个词语的 TF-IDF 值均值，作为该句评论的特征进入模型。

表 4.2 样例评论中每个词语的 TF-IDF 值

鞋子	很	好	总体	上	穿着	还	算	舒服
0.074	0.064	0.084	0.186	0.137	0.098	0.077	0.205	0.088
物流	足够	快	唯一	遗憾	左脚	脚后跟	一点点	磨脚
0.195	0.256	0.205	0.256	0.233	0.233	0.217	0.256	0.233

(2)评论长度：数据来源于实际生活，存在评论长短不一致的现象，若评论文本长度较长，包含的信息量较多，表述内容就越详尽，可以吸引更多潜在顾客阅读浏览。而较短的文本获取的信息量较少，因此在特征中加入原始评论全部的字符数作为评论长度。

(3)是否包含煽动性词语：有学者研究发现[43]，虚假评论中经常含有煽动性的词语，如“别再犹豫，赶紧下单”，评论中带有诱导性质的词语，这个特征能够较好的反映虚假评论的一些专有特性。

(4)是否提及电商平台名称：该特征用于判断评论中是否提交电商平台，若评论中提及电商平台名称而非着重描述商品本身，则是虚假评论的概率较高。

(5)是否提及品牌名称：虚假评论发表者由于任务时间限制等原因不会阅读商品的详细介绍，仅用通用性词汇进行评论，如只介绍快递情况、与品牌商品无关内容等。

(6)评论包含图片数量：一般发表虚假评论会带有多张图片，来增加评论内容的说服力，以增加商品评论的可信度，而真实顾客的评论少有评论带有图片。

(7)追评数量/回复追评数量：刷单团队在一定时间段内进行某项刷单任务，完成指定商品评论内容，不会在后续回复其他顾客对于评论提出的问题，这些特征也能较好的区分虚假评论和真实评论。

4.2 模型评价指标

对于利用样本训练得到的模型，需要进行模型评估，体现模型性能。利用 PU 学习算法来识别虚假评论本质是一个文本二值分类问题。对于二分类问题，预测结果可能出现四种情况：真正类(true positive，TP)，属于正例的样本被正确预测为正；假正类(false positive，FP)，属于正例的样本被正确预测为负；真负类(true negative，TN)，属于负例的样本被正确预测为负；假负类(false negative，FN)，属于负例的样本被正确预测为正，用混淆矩阵展示四类结果，结果如表 4.1 所示：

表 4.3 混淆矩阵

	预测为正	预测为负
真实为正	真正类(true positive，TP)	假负类(false negative，FN)
真实为负	假正类(false positive，FP)	真负类(true negative，TN)

目前模型常用的评价指标有准确率(Accuracy)、精确率(Precision)、召回率(Recall)和 F1 值[44]。准确率(Accuracy)为预测正确的正例和负例样本数占全部样本的比例，精确率(Precision)为预测为正例的样本中预测正确的比例，召回率(Recall)为正例样本中被预测正确的比例，F1 值是精确率和召回率的调和平均数。

与监督学习不同的是，本文利用半监督学习所研究的文本分类的类别分类严重不平衡，往往精准率和召回率是相互矛盾的度量标准，一般情况下精确率(Precision)高的时候，召回率(Recall)较低，当召回率(Recall)高时，精准率(Precision)较低。参考前人[45]的经验，选用 F1 值度量分类器效果，F1 值具体计算公式如下所示：

$$F1 = \frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

4.3 分类器选择

本文 PU 学习算法包括两步，第一步和第二步均使用分类器进行学习，第二步对未标记样本进行分类直至迭代结束。第 2 章 2.2 节介绍的朴素贝叶斯、支持向量机、fastText、GBDT、XGBoost、LightGBM 共 6 种分类器适用场景不同，各有优缺点，因此选用 6 种分类器与 PU 学习算法相结合，对比使用不同分类器对虚假评论识别模型效果差异。分类问题中，监督学习和半监督学习都是对已标记的样本进行学习以达到分类的目的。当模型中标记的数据量越大时，模型学习到的数据信息和特征越完善，分类效果越好。因此本文随机从已标记的正例商品评论中随机抽取 300、500 条正例样本，并使用标记的负例样本共 1000 条数据进入模型进行学习分类。

为保持模型对比的有效性，除了分类器和正例样本数量不同，训练数据预处理方法和过程均保持一致。

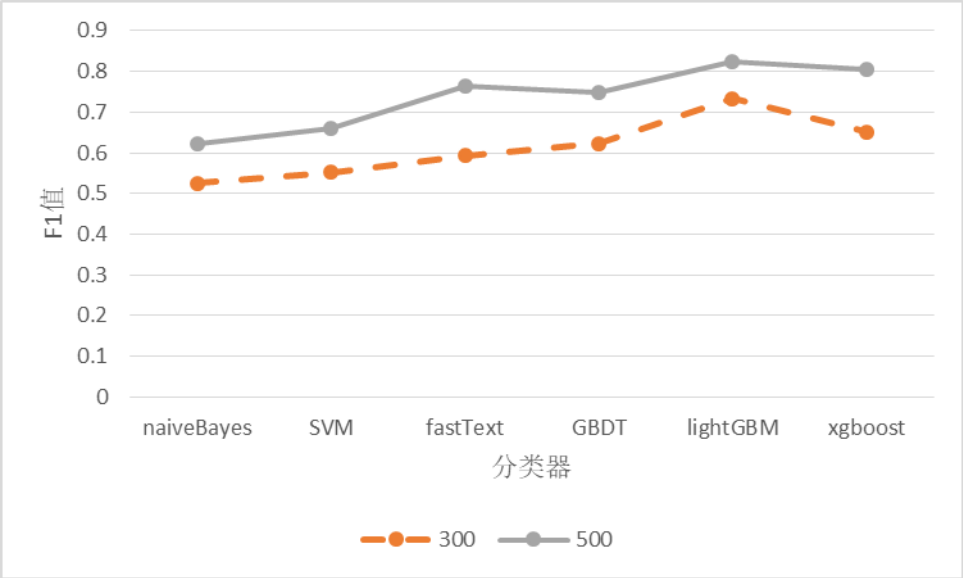

图 4.2 不同正例样本数量的分类效果

图 4.2 展示不同比例的正例样本下，各个分类器模型分类的效果，当进入模型的正例样本数量越多的时候，分类器的效果越好。

表 4.4 模型效果的 F1 值

正例样本数量	300	500
naiveBayes	0.525	0.623
SVM	0.552	0.660
fastText	0.593	0.764
GBDT	0.622	0.749
LightGBM	**0.734**	**0.824**
XGboost	0.651	0.805

表 4.4 展示应用各个分类器和不同正例样本比例的模型效果，随着正例样本数量的增加，模型效果随之变好，且 LightGBM 分类器的分类效果均优于其他分类器。当使用 500 个正例样本，PU 学习算法与 LightGBM 分类器相结合时，F1 值达到 0.824，分类效果较好，能够较好的区别虚假评论和真实评论。

4.4 本章小结

本章节主要阐述特征构建、模型评价指标选取和分类器选择。首先介绍特征工程，从电商平台获取的原始评论格式较为混乱，且计算机不能直接识别其特征，需要从原始

评论中提取关键特征。其次介绍模型评论指标，分析现有分类评价指标后选用 F1 值，对比各个模型之间的效果。最后进行模型训练，使用 PU 学习算法结合 6 种分类器和不同数量的正例样本进行模型训练，对比各个模型之间的分类效果，选用效果最好的 LightGBM 算法与 PU 学习算法结合，进行电商平台真实商品的虚假评论识别体系的构建。

5 模型应用

本章首先对上一章节的模型进行实例分析，分析的数据采用某品牌鞋子的评论数据，并对预测结果进行可视化展示和特征分析，根据分析结果为消费者提供相关参考意见。其次将构建的虚假评论识别体系进行应用商品拓展，可完成他类商品虚假评论的识别。

5.1 应用预测

上文已确定使用 PU 学习算法和 LightGBM 结合能够较好的识别虚假评论，并将模型应用于全部数据集进行预测，预测评论数据所属标识(虚假评论或真实评论)。为更好地展示两类评论之间的差异，本节利用虚假识别模型对已识别的虚假评论数据进行分析。主要包括结果可视化和特征分析，分析虚假评论和真实评论之间的差异，并为消费者简便区分虚假评论与真实评论的建议。

5.1.1 高频词汇展示

文本挖掘中经常使用词云图，词云图众多文本中出现频率较高的词语在视觉上得以突出，且能够快速直观地将评论中的关键信息直接出来，每个词语的重要性通过字体大小得以显示，字体越大，在图片中越突出，这个词语就越重要。本文通过对高频词汇生成词云，可视化展示商品评论的高频词汇，绘制词云图可以快速得知评论主体表达的意思，迅速抓住重点。

根据模型识别的虚假评论和真实评论，为避免敏感信息，剔除关于电商平台和商品品牌名称，两类评论分别随机抽取 200 条的评论进行可视化展示。利用词云图对比虚假评论和真实评论的高频词汇之间的差异，图 5.1 展示预测标记为 1，即虚假评论的词云图，图 5.2 展示预测标记为 0，即真实评论的词云图：

图 5.1 鞋子虚假评论的词云图

图 5.2 鞋子真实评论的词云图

从虚假评论的词云图中可以看出，出现频率较高的词汇为：“鞋子”、“非常”、 “不错”、“质量”、“舒服”、 “喜欢”，“物流”、“快递”等。而真实评论的词云图中出现频率较高的词汇为：“鞋子”、“不错”、“舒服”、“有点”、“舒服”、 “喜欢”、“物流”等。

除去“鞋子”、“不错”、“舒服”、“喜欢”等在两类评论均出现的高频率词语，对比两类评论之间的差异只要有以下两点：

(1)情感基调不同：虚假评论中表达的情感基调较为饱满，情绪高昂。如“非常”、 “特别”这样的词语出现的频率较高，而在真实评论中表现的情感较为婉转，如“感觉”、“有点”均表达的是较为收敛的情绪。

(2)表达内容不同：虚假评论中描述的多为快递方面，“快递”、“物流”两个出现的频率很高。由于刷单团队没有收到商品和真实试穿，没有具体的细节描述，只能从物流快递角度出发撰写评论。对于真实评论而言，真实购买的顾客会从各个方面出发描述收到的商品和体验感受，因此各特征词较为分散，在词语图中未明显显示。

为了展现词云图的高频词汇出现的具体频词，经过统计汇总展示频词前 10 的词语，具体如表 5.1 所示：

表 5.1 两类评论高频词汇统计

虚假评论		真实评论	
词语	频数	词语	频数
鞋子	122	鞋子	60
不错	91	不错	54
质量	73	舒服	42
舒服	67	穿着	34
非常	56	感觉	31
收到	51	有点	30
穿着	46	可以	27
物流	45	喜欢	19
快递	45	质量	19
喜欢	41	鞋底	19
购买	37	价格	19
值得	32	透气	19
购物	31	好看	19
满意	31	起来	19
一次	29	没有	19

表 5.1 发现，即使取相同的评论数，虚假评论和真实评论出现频词最高的两个词语均为“鞋子”、 “不错”，但是虚假评论中“鞋子”的频数是 122，而真实评论全部评论中仅有 60 个词语为“鞋子”，频数之间相差较为悬殊。在此基础上，计算两类评论的平均句子长度，虚假评论评论每句评论有 55 个字符，而真实评论仅是 27 个字符，即虚假评论的评论文本长度较长，词语出现的频数总体较高。

5.1.2 特征对比分析

上一小节通过评论内容高频词汇的可视化展示，并进行高频词频的统计，从整体上展现两类评论之间的差异性。本节利用 TF-IDF 方法提取两类评论之间的主要特征词进

行对比分析。

TF-IDF 结合词频和逆文本频率两种方法，从词语出现在该评论中的频率和总体分布情况两个方面来衡量词语的重要性。与词云图不同，TF-IDF 需要结合词语在当前评论中的分布，还考虑词语在全部评论中的分布，若一个词语在当前评论中出现的频率较高，在全部评论中出现的频率较低，那么这个词语的代表性较好，因此 TF-IDF 值能较好的抽取每句评论中的文本特征。

为与 5.1 节词云图方便对比，数据预处理与 5.1 节保持一致，汇总 200 条虚假评论和真实评论，将每句评论进行分词，去除停用词，语料库构建，向量转换，运用公式(2.1)计算每个词语的 TF-IDF 值，并对全部评论中每个词语 TF-IDF 值求均值，以近似表示该词语在虚假评论或真实评论中整体重要性，得到虚假评论中 TF-IDF 值最高的前 10 个特征词，如表 5.2 所示。

表 5.2 虚假评论主要特征词及其平均 TF-IDF 值

特征词	平均 TF-IDF 值
实用	0.768
美观	0.768
完美	0.461
可以信赖	0.384
价廉物美	0.354
优惠活动	0.354
大气	0.354
最舒服	0.329
大方	0.328
称赞	0.307

根据相同的方法处理数据，计算真实评论各个特征的 TF-IDF 值，取 top10 的特征词，如表 5.3 所示。

表 5.3 真实评论主要特征词及其平均 TF-IDF 值

特征词	平均 TF-IDF 值
太轻	4.605
真不错	1.151
脱胶	1.399
很漂亮	1.151
一般	0.853
开胶	0.839
一般般	0.815
颜值	0.807
偏硬些	0.768
太薄	0.768

从主要特征词角度分析发现，虚假评论与真实评论的内容之间主要存在以下两个方面差异：

(1) 情感倾向方面：虚假评论的主要特征词中均是褒义词和赞美词，如“完美”、“称赞”等词高度夸赞商品，而真实评论中通过购买者真实体验发现鞋子仅一般，并未像虚假评论评论的那般完美。

(2) 质量问题方面：虚假评论描述鞋子“最舒服”、“物美价廉”，欲增加潜在购买者的购买决心，而通过真实评论却反映鞋子“太薄”、“偏硬些”且存在“开胶”的现象，存在一定的不良体验。

5.1.3 消费者建议

通过模型应用发现虚假评论内容多涉及物流、快递等，评论内容与商品特征关联性低，较为宽泛，并存在过度夸赞的情况。而真实评论内容涉及商品的价格、舒适度等，且真实评论的 TF-IDF 值高于虚假评论，说明模型具有较好的可识别性。

应用本文构建的虚假评论识别体系，将鞋子的网购评论区分为虚假评论和真实评论，通过分析虚假评论和真实评论的高频词汇和特征词汇，发现虚假评论会高度夸奖商品，从中表达的情绪较为高昂，常出现“非常”、“特别”等词汇，且较多表达对快递和物流的看法。真实评论表达商品真实的用户体验，描述商品的细节、质量、缺点等多个方面。

因此对于消费者而言，在选购商品期间，可参考上述对虚假评论和真实评论的差异，若商品评论中出现“非常”、“特别”等情绪高昂的词汇，从物流角度描述且评论内容与商

品特征关联性低，较为宽泛，该评论为虚假评论的可能性较大；若商品评论从多个角度描述商品质量、用后体验等，或指出商品一些缺点，该评论为真实评论的可能性较大。消费者参考买家的真实评论后，从商品评论中获取更有价值的信息，挑选心仪的商品。

5.2 应用商品拓展

本文构建的虚假评论体系能够快速有效的识别虚假评论，可以较好的识别虚假评论和真实评论，且在应用价值上具有一定的可拓展性，可应用于不同类型的商品评论，进行各种商品虚假评论的识别。

本节将上述构建的虚假评论体系应用于裤子，研究某电商平台获取某品牌的裤子，获取评论 1850 条，进行数据清洗，标记 200 条正例样本和 200 条负例样本，使用 PU 学习算法和 LightGBM 识别虚假评论，在 400 条标记样本中模型的 F1 值达到 0.807，分类效果较好，反映模型能较好的识别其他品类商品的虚假评论，具有良好的拓展性。

模型应用于全部评论数据集进行预测，从预测结果的两类评论中分别随机抽取 150 条的评论进行词云图绘制。使用词云图对比裤子的两类评论之间的差异，图 5.3 展示虚假评论的词云图，图 5.4 展示真实评论的词云图：

图 5.3 裤子虚假评论的词云图

图 5.4 裤子真实评论的词云图

根据图 5.3 和图 5.4，分析裤子的虚假评论与真实评论的高频词汇，除去两类评论中都出现的词语，虚假评论中“非常”、“特别”、“好评”、“满意”、“快递”等词语出现的次数较多，评论表达的情感基调较为饱满，商品的评价较高，多从快递角度进行表述。而真实评论中的高频词汇为“有点”、“可以”、“感觉”、“面料”、“版型”、“布料”、“弹性”，表达的情绪较为中肯，从裤子的多个角度进行商品论述。

利用 5.1.3 节的 TF-IDF 计算方法，提取两类评论之间的主要特征词进行对比分析，表 5.3 展示裤子虚假评论主要特征词及其平均 TF-IDF 值，表 5.4 展示裤子虚假评论主要特征词及其平均 TF-IDF 值：

表 5.3 裤子虚假评论主要特征词及其平均 TF-IDF 值

特征词	平均 TF-IDF 值
物美价廉	3.401
完美	1.956
太好了	1.304
第二条	1.304
合身	1.263
一如既往	1.119
最佳	1.079
物超所值	1.079
很正	1.079
大牌	0.869

表 5.4 裤子真实评论主要特征词及其平均 TF-IDF 值

特征词	平均 TF-IDF 值
低腰	1.812
好看	1.701
太薄	1.701
裤料	1.304
弹性	1.119
挺大	1.079
弹力	1.022
贴身	0.978
挺舒服	0.978
粘灰	0.864

从 TF-IDF 值角度分析，裤子虚假评论的主要特征词中多是赞美词，如“价廉物美”、“物超所值”、“完美”、“最佳”等词语高度夸赞商品，且词语较为概括，而真实评论的用户已收到商品，评论内容从裤子的裤型、布料、穿着体验等细节描述真实用户感受。

5.3 本章小结

本章节阐述应用预测和应用商品拓展两个方面。应用预测从高频词汇的词云展示和重要特征词的提取，从整体和具体两个方面展开分析虚假评论与真实评论之间的差异。利用 python 绘制高频词汇的词云图，发现真实评论和虚假评论的词云图和词频均存在明显差异。进一步利用模型对真实评论和虚假评论的 TF-IDF 值进行具体的分析，真实评论的 TF-IDF 值高于虚假评论，说明虚假评论的主体内容较为相似，而真实评论反映顾客的真实体验，从而给予消费者区别虚假评论的相关建议。应用商品拓展将本文构建的虚假评论体系应用于其他商品，获取电商平台上裤子的评论数据，经过数据清洗、数据标记、模型构建与应用，进行虚假评论识别，能够较好的区分虚假评论与真实评论，验证本文模型具有较好的可扩展性。

6 结论与展望

6.1 结论

电子商务的崛起反映了大量商家选择通过开网店的方式来代替传统店铺经营模型。网店评论作为消费者购物的一个重要依据，成为商家和买家关注的焦点。部分商家选择利用刷单团队进行虚假好评提高店铺好评率，从而误导消费者购物。如何识别虚假评论对消费者和监管部门具有重要的实际意义电子商务的崛起反映了大量商家选择通过开网店的方式来代替传统店铺经营模型。网店评论作为消费者购物的一个重要依据，成为商家和买家关注的焦点。部分商家选择利用刷的团队进行虚假好评提高店铺好评率，从而误导消费者购物。如何识别虚假评论对消费者和监管部门具有重要的实际意义。

本文探索了虚假评论领域众多学者的研究成果，针对评论内容方面的研究主要分为三大类：监督学习、无监督学习、半监督学习，为了提高模型准确性和减少数据标注的工作，本文选用半监督学习领域的 PU 学习算法来识别虚假评论，并预测得到的虚假评论进行特征分析。本文研究主要取得如下成果：

(1)基于半监督学习算法完成一套完善的虚假评论识别体系，使用 PU 学习算法与朴素贝叶斯、SVM、fastText、GBDT、LightGBM、XGBoost 算法结合，选取识别效果最好的分类器与PU学习算法结合，能够准确快速地识别虚假评论。

(2)构建中文虚假评论标记数据集，从现实生活中实时爬取店铺评论数据源，获取数据源之后，利用 python 进行数据预处理，剔除特俗字符、数字、标点等无关信息。利用筛选重复评论和时间序列异常的数据，使用机器学习和专家标注的方法构建标记数据集，减少人工标注的主观臆断，从而提高标注效率。

(3)在应用模型识别虚假评论之后，从整体高频词汇和具体特征分析虚假评论的特征，分析发现与现有学者的研究成果基本一致，并将模型应用于其他商品评论，说明本文的虚假评论识别体系具有良好的可拓展性。

本文完成数据源获取、数据预处理、特征提取、模型构建、虚假评论特征分析，搭建一套虚假评论识别的体系，能够较好的区分虚假评论和真实评论，可推广使用至其他电商平台和其他网购商品的虚假评论识别。

6.2 展望

对于本文研究的虚假评论识别体系有以下两点展望，在未来可以展开进一步的研究：

1、在识别虚假评论中，使用的数据维度对于模型的准确性和精度有一定的影响，本文主要从网购评论的文本内容出发，获取的数据特征较少，因此应该分析更丰富的特征以更精确地识别虚假评论。本文采集了电商平台上商品的评论文本、评论是否包含图片、是否追评等维度，未来还可以分析评论带有的图片内容是否与购买商品保持一致；从使用评论者角度获取特征，如同一个评论者在不同商品下的评论内容与图片的相似度等，获取并分析更有意义的特征，能够更好的识别虚假评论。

2、由于PU学习算法是迭代算法，在第二步的迭代过程中会有噪音数据，影响了分类的精度，后续可以从算法优化的角度研究避免噪音的方法，降低噪音数据对分类的影响，以达到提高模型识别精度的效果。

参考文献

[1] 李存林，杨世瀚，王晗. 基于隐含语义分析的电商虚假评论识别[J]. 广西民族大学学报(自然科学版). 2018, 24(1): 52-59.

[2] Li F, Huang M, Yi Y, et al. Learning to Identify Review Spam[C]. Twenty-second International Joint Conference on Artificial Intelligence. 2011:2488-2493.

[3] Yazdanifard R, Simon K D, Ming C W, et al. The Impact of Online Consumer Reviews on Sales in B2C E-Commerce[J]. 1973, 57(1): 341-349.

[4] Ullrich S, Brunner C B. Negative Online Consumer Reviews: Effects of Different Responses[J]. Journal of Product & Brand Management. 2015, 24(1): 66-77.

[5] Jindal N, Liu B. Review Spam Detection [C]. In: Proceedings of the 16th International Conference on World Wide Web (WWW'07). New York: ACM, 2007: 1189-1190.

[6] 朱娟. 在线商品虚假评论关键问题研究综述[J]. 现代情报. 2017, 37(5): 166-171.

[7] Jindal N, Liu B. Opinion Spam and Analysis[C]. ACM. 2008: 219-230.

[8] Li J, Ott M, Cardie C, et al. Towards a General Rule for Identifying Deceptive Opinion Spam[Z]. Baltimore, Maryland: 2014:1566-1576.

[9] Li H, Chen Z, Liu B, et al. Spotting Fake Reviews via Collective Positive-unlabeled Learning[C]. 2014 IEEE international conference on data mining. IEEE, 2014: 899-904.

[10] Lau R Y, Liao S Y, Kwok R C, et al. Text Mining and Probabilistic Language Modeling for Online Review Spam Detection[J]. ACM Transactions on Management Information Systems (TMIS). 2011, 2(4): 1-30.

[11] 宋海霞，严馨，余正涛，等. 基于自适应聚类的虚假评论检测[J]. 南京大学学报(自然科学版). 2013, 49(4): 433-438.

[12] Zhu X, Goldberg A B. Introduction to Semi-supervised Learning[J]. Synthesis Lectures on Artificial Intelligence and Machine Learning, 2009, 3(1): 1-130.

[13] Liu B, Dai Y, Li X, et al. Building Text Classifiers Using Positive and Unlabeled Examples[C]. Third IEEE International Conference on Data Mining. IEEE, 2003: 179-186.

[14] Hern A Ndez Fusilier D, Guzm A N Cabrera R, Montes-Y-G O Mez M, et al. Using PU-Learning to Detect Deceptive Opinion Spam[Z]. Atlanta, Georgia, 2013:38-45.

[15] Ott M, Choi Y, Cardie C, et al. Finding Deceptive Opinion Spam by any Stretch of the

Imagination[C]. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-volume 1,2011: 309-319.

[16] 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[J]. 计算机研究与发展. 2015, 52(3): 639-648.

[17] Lim E P, Nguyen V A, Jindal N, et al. Detecting Product Review Spammers Using Rating Behaviors[C]. Proceedings of the 19th ACM International Conference on Information and Knowledge Management. ACM, 2010: 939-948.

[18] Wang G, Xie S, Liu B, et al. Review Graph Based Online Store Review Spammer Detection[C] IEEE 11th International Conference on Data Mining. IEEE, 2011: 1242-1247.

[19] 富越，董保华，Fuyue，等. 电子商务垃圾评论者识别研究[J]. 科学决策. 2015(9): 79-94.

[20] 吴佳芬，马费成. 产品虚假评论文本识别方法研究述评[J]. 数据分析与知识发现. 2019, 3(9): 1-15.

[21] 陈燕方. 基于DDAG-SVM的在线商品评论可信度分类模型[J]. 情报理论与实践. 2017, 40(7): 132-137.

[22] 王梦华. 基于半监督学习的虚假评论识别研究[D]. 南京财经大学, 2018.

[23] 王禹. 电商平台购物虚假评论识别研究[D]. 首都经济贸易大学, 2018.

[24] Li H, Liu B, Mukherjee A, et al. Spotting Fake Reviews Using Positive-unlabeled Learning[J]. Computaciony Sistemas, 2014, 18(3): 467-475.

[25] 张荷，李梅，张阳，等. 基于PU学习的软件故障检测研究[J]. 计算机应用研究. 2015, 32(11): 3324-3327.

[26] 王艺琪. 基于PU学习的磷酸激酶抑制剂筛选算法[J]. 信息通信. 2016(7): 53-55.

[27] 隋福宁，杨强. 一种基于改进PU学习理论的推送内容过滤策略[J]. 计算机应用研究. 2010, 27(12): 4480-4482.

[28] Fusilier D H A N, Montes-Y-G O Mez M, Rosso P, et al. Detecting Positive and Negative Deceptive Opinions Using PU-learning[J]. Information Processing & Management. 2015, 51(4): 433-443.

[29] Dietterich T G. Ensemble Methods in Machine Learning[C]. Springer, 2000:1-15.

[30] Breiman L. Bagging Predictors[J]. Machine Learning. 1996, 24(2): 123-140.

[31] Johnson R W. An Introduction to the Bootstrap[J]. Teaching Statistics. 2001, 23(2):

49-54.

[32] Schapire R E. The Strength of Weak Learnability[J]. Machine Learning. 1990, 5(2): 197-227.

[33] Friedman J H. Greedy Function Approximation: A Gradient Boosting Machine[J]. Annals of Statistics, 2001: 1189-1232.

[34] Chen T, Guestrin C. Xgboost: A Scalable Tree Boosting System[C]. ACM, 2016:785-794.

[35] 李叶紫,王振友,周怡璐,韩晓卓.基于贝叶斯最优化的Xgboost算法的改进及应用[J]. 广东工业大学学报,2018,35(1):23-28.

[36] Ke G, Meng Q, Finley T, et al. Lightgbm: A Highly Efficient Gradient Boosting Decision Tree[C]. Advances in Neural Information Processing Systems. 2017: 3146-3154.

[37]李丹. 基于朴素贝叶斯方法的中文文本分类研究[D].河北大学,2011.

[38] Yu C T, Salton G. Precision Weighting---An Effective Automatic Indexing Method[J]. Journal of the Acm.1975, 23(1): 76-88.

[39] Amati G, Van Rijsbergen C J. Probabilistic Models of Information Retrieval Based on Measuring the Divergence from Randomness[J]. ACM Transactions on Information Systems (TOIS), 2002, 20(4): 357-389.

[40] 武永亮，赵书良，李长镜，等. 基于TF-IDF和余弦相似度的文本分类方法[J]. 中文信息学报. 2017, 31(5): 138-145.

[41] 王军，李鑫. 网络评论信息对消费者购买态度的影响研究[J]. 情报理论与实践. 2014, 37(9): 121-124.

[42] Narayan R, Rout J K, Jena S K. Review Spam Detection Using Semi-supervised Technique[C]. Progress in Intelligent Computing Techniques: Theory, Practice, and Applications, Springer, 2018, 281-286.

[43] 沈超，刘士伟，徐滔. 电商平台商家诱导评论的特征与对策研究[J]. 电子商务. 2019(5): 47-49.

[44] Powers D M. Evaluation: from Precision, Recall and F-measure to ROC, informedness, markedness and correlation[J]. Journal of Machine Learning Technologies, 2011, 2(1), 37-63.

[45] 郭芷榕，王会青，白莹莹. 基于PU学习和自主训练的时间序列分类模型[J]. 计算机工程与设计. 2018, 39(9): 2780-2786.

致谢

光阴似箭，日月如梭，三年的硕士时光转眼即逝，马上就进入尾声，刚入学的情景仿佛就在眼前，回顾在研究生生涯的学习收获颇多，感谢让我有幸师从黄恒君导师门下，感恩遇到的点滴。

首先感谢我的导师黄恒君教授，黄老师拥有渊博的学识，严谨的工作态度、谦逊的人格魅力，是我学习的榜样。在校期间，黄老师在学习和生活上对我的谆谆教诲，并给予我悉心的指导和帮助。研一刚入学的时候尽心帮助规划三年的学习方向，在每周讨论班中收获前沿学术成果，解决学术和学习上的疑惑，在毕业论文的选题方向、撰写过程到后续修改，都离不开黄老师的指导。

其次感谢许腾腾师兄，在遇到学术困难时给予的帮助，在学习过程中给予的鼓励；感谢刘云师兄分享就业的众多感悟；感谢同门卢旺同学分享学习和生活中的苦与乐；感谢室友杨朝雯和张亚凡，当我孤身一人在异乡求学时给予的帮助和温暖。感谢在研究生生涯中遇到的你们，陪伴我度过颇有收获的三年研究生时光。

感谢父母与家人，对我的付出和支持，是你们多年来给我的鼓励和理解才使我能够积极快乐的成长，让我有机会得以学习更多的知识。

最后向百忙之中评审本论文的老师和进行答辩指导的各位老师表达真诚的感谢！