分类号 C8/284　　密级

U D C　　编号 10741

兰州财经大学

LANZHOU UNIVERSITY OF FINANCE AND ECONOMICS

硕士学位论文

（专业学位）

论文题目　基于文本挖掘的虚假评论识别

研究生姓名：倪志恒

指导教师姓名、职称：杨盛菁 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：2021 年 6 月 6 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：_____ 签字日期：2021.6.6

导师签名：_____ 签字日期：2021.6.6

导师(校外)签名：_____ 签字日期：2021.6.6

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：_____ 签字日期：2021.6.6

导师签名：_____ 签字日期：2021.6.6

导师(校外)签名：_____ 签字日期：2021.6.6

False comment recognition based on Text Mining

Candidate : Ni Zhiheng

Supervisor: Yang Shengjing

摘 要

近年来，随着新零售以及移动网络支付的高速发展，人们对于消费途径的多样性需求越来越旺盛，网上购物成为人们消费、购物的重要途径。与此同时，网上购物会产生很多的在线评论，这些评价对商家的销量产生很大的影响，因为互联网具有高度开放性的特点，有很多的商家开始关注消费者网络购物产生的网上评论信息，在利益的驱动下，出现了商家背后操纵评论的现象，网上开始出现虚假评论，严重损害了消费者的权益。因此，无论是从消费者个人的角度出发还是从商家、平台的角度出发，对虚假评论信息进行识别都是一项急需解决的工作。但是在线评论每日的增量都很巨大，利用人工对评论进行审核会耗费大量的人力物力，所以需要一套有效的识别方法对在线评论进行自动甄别，有效剔除虚假评论。

本文的目的是能够给出一套针对电商平台的虚假评论识别方法，能够快速、精准、有效地识别虚假评论，并总结出虚假评论的行为模式。本文对虚假评论识别的主要工作包括：

首先从电商平台获取相关的评论数据，对文本数据进行清洗，将分好词的文本数据进行合适的向量化操作，并通过评论发布时间、重复评论、评论者情感倾向等对数据进行标注。

其次，将向量化后的文本采用传统的机器学习和深度学习等方法进行基分类器的训练，并通过集成学习方法将各个基分类器进行组合，对文本评论数据进行识别。

利用训练好的模型对大量的未标注评论数据进行预测，通过虚假评论的语言模型、分析主题差异、进行词性分析、情感分析等工作挖掘虚假评论的行为属性及其行为模式。实验结果表明：本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，具体一定的实际应用价值。

关键词：虚假评论 文本挖掘 神经网络 模型融合 行为模式

Abstract

In recent years, with the rapid development of new retail and mobile network payment, people's demand for the diversity of consumption channels is more and more vigorous, and online shopping has become an important way for people to consume and shop. At the same time, online shopping is producing a lot of online comments, the evaluation of a large impact on sales to the businessman, because the Internet has the characteristics of highly open, there are a lot of businesses begin to pay close attention to the consumer online shopping, online comments information generated under the drive of interests, the business behind the phenomenon of comments, began to appear false comments online, serious harm the rights of consumers. Therefore, whether from the perspective of individual consumers or from the perspective of businesses and platforms, it is an urgent task to identify fake review information. However, the daily increment of online comments is huge, and it will cost a lot of manpower and material resources to review comments manually. Therefore, a set of effective identification method is needed to automatically screen online comments and effectively eliminate false comments.

The purpose of this thesis is to provide a set of identification methods for fake comments on e-commerce platforms, to identify fake comments quickly, accurately and effectively, and to summarize the behavior patterns of fake comments. This thesis mainly realizes the identification of false

comments based on the method of text mining. The main work includes:

Firstly, the relevant comment data were obtained from the e-commerce platform, and the text data were cleaned. The text data with good words were subjected to appropriate vectorization operation, and the data were annotated by comment release time, repeated comments, and the emotional tendency of the commenters.

Secondly, the vectorized text is trained with the traditional machine learning and deep learning methods, and each base classifier is combined with the ensemble learning method to recognize the text comment data.

The trained model is used to predict the data of a large number of unlabeled comments, and the behavior attributes and behavior patterns of fake comments are mined through language model of fake comments, topic difference analysis, part of speech analysis, sentiment analysis and other work. The experimental results show that this method has a good performance in the identification of false comments, which provides a new method for consumers and regulators, and has a certain practical application value.

Keywords：False comments; Text mining; Neural network; Model fusion; Behavior pattern

目 录

1.引言

1.1 问题的提出及研究意义

1.1.1 问题的提出

近年来，在我国经济平稳增长以及移动支付技术不断更新、发展的时代背景下，我国的网络购物用户呈现出不断增长的态势，电子商务市场蓬勃发展，与此同时，消费者们逐渐养成了网络购物的消费习惯，网络零售持续发展，成为我国经济增长的重要动力。根据第 45 次 CNNIC 的数据显示，截止到 2020 年 3 月，我国网络购物者的规模不断扩大，达到了 7.49 亿，约占据网民总数的 79%；与此同时，手机网络购物的用户规模也占据了很大的比例，达到了 7.47 亿，占据全体手机网民的 80%，电子商务、网络消费等通过各种新模式，不断释放出其动能，从各个方面拉动我国的经济增长[1]。社交电商、直播电商成为网络消费增长的新功能，消费潜能能够进一步释放，推动网络零售的发展。并可以预计，未来几年 B2C 网络销售市场的规模将与 C2C 市场销售规模进一步拉开差距，并在电子商务领域处于主导地位。

在网络零售业以及电子商务蓬勃发展的背景下，诞生了各种各样的移动网络应用、网上商城，在方便人们日常生活、购物的同时，也提供了一种新的消费者表达自己想法的方式，消费者们可以在这些平台上面，充分表达自己的观点、对商品的使用体验、看法，以作为参考，在这些网络平台之上出现了大量的消费者对于商品的在线评论。这些评论是消费者了解想要购买的商品信息的宝贵来源，消费者们可以通过这些在线评论来了解自己想要购买的商品，这些评论也会成为消费者是否购买该产品的决定性因素，消费则购物之前一般都会查看商品的在线评论以作参考，因此这些评论会直接和产品的销量相挂钩。

根据相关的研究我们可以发现，绝大部分的消费者都会相信网上的线上评论信息，而且积极的在线评论会使得消费者会更加的相信商家，增加商家的销量[17]。在另一方面，平台上存在的商品的在线评论信息对于商家来说，也有指出其商品不足之处，以及帮助其提升商品的质量、改善服务。由此可以看到，各个平台上

存在的在线评论无论对于商家还是消费者来说都具有者很大的价值。

因为互联网具有高度开放性的特点，有很多的商家开始关注消费者网络购物产生的网上评论信息，但是由于任何人都可以低成本的在平台上发布自己的意见和评论，甚至出现了商家之间的恶性竞争，开始操纵这些评论信息。有些商家会通过雇佣某些组织或者人员来给自己的商品撰写好评，还有甚者会给竞争商家发布消极评论，诋毁商品。这种虚假评论现象的出现，严重损害了消费者的权益。这些评论丧失了其本身所具有的价值和意义。

因此，无论是从消费者的角度出发还是从商家、平台的角度出发，对虚假评论信息进行识别都是一项急需解决的工作。网上评论每日的增量巨大，而且虚假评论具有伪装性和欺骗性的特点，利用人工对评论进行审核会耗费大量的人力物力，所以需要一套有效的识别方法对评价评论进行自动甄别，有效剔除虚假评论，给商家营造出一个公平的网络竞争环境，同时维护消费者的权益。

1.1.2 研究目的和意义

在我们的日常生活当中，非结构化数据也扮演着非常重要的角色，本文基于网上商城的评论文本进行相关研究，也属于非结构化数据的建模分析。本文研究的目的在于能够根据评论本文进行建模、学习，以此建议一套能够针对该领域的大规模文本快速、准确地识别其是否为虚假评论的方法，并根据其分类结果对大规模评论语料进行识别，分析其隐含的行为特征以及语言特征。主要目标有两个：一是模型准确、稳定性好。在现在的电商平台当中，评论语料每天都在大量的增加，完全根据人力对评论语料进行人工审核，工作量十分巨大，会耗费大量的人力、物力。本文采用传统的机器学习和神经网络深度学习相结合的方法，对文本语料进行学习，并采用集成学习的方法提升模型的稳定性、准确性。二是利用训练好的模型对大量的未知数据进行预测，对虚假评论进行识别，对通过大量的评论数据进行对比、建模分析其存在的差异，深入地发掘虚假评论隐藏的语言特征和行为特征。

在理论意义上，本文的研究主要针对电商平台的评论语料进行建模，进行虚假评论的识别，为我国该领域的研究提供一定程度上的参考。目前，对于虚假评论的研究主要集中在国外，学者们利用亚马逊数据集或其他的数据集进行虚假评

论的识别研究，但是因为中英文的语言差异，这些研究不能够直接对中文语料进行分析。在国内，学者对于虚假评论的研究还比较少，大部分都使用某种基分类器进行文本评论的识别，稳定性不是很好，国内对于文本评论这种非结构化数据的研究主要集中于评论的情感分析以及评论语料的有用性研究等，所以本文的研究可以为国内该领域的研究提供某种程度上的参考，具有一定的理论意义。

在现实意义上，通过训练好的模型对在线评论进行虚假评论的识别并对虚假评论分析其语言特征和行为特征，能够有效的对虚假评论进行大规模化的识别。可以知道，对于电商平台来说，随着商品的售出，针对商品的在线评论也在不断地增加着，某些大型的电商平台，评论的日增量无疑更大，要对这些评论一一人工识别，这种耗费人力物力的方式是不大可能实现的，但是通过模型去迭代训练就可以很好的解决这个问题，不仅能够规模化的识别，可以实现低成本且准确的识别虚假评论，提升工作效率。

本文通过文本挖掘的方式对虚假评论进行识别，随后利用 lda 主题聚类、词性分析以及统计语言模型等方法对虚假评论和真实评论的特点进行对比分析，分析虚假评论的语言特征和行为特征，尤其是针对真实评论里面消费者所表达的满意点和不满的地方，可以为商家改进自身服务、提升商品质量提供帮助。

另外，通过大量的在线评论的识别操作，可以统计出某个平台的在线评论质量状况，为整个平台的商品质量、用户活跃状况提供参考，帮助平台净化环境。此外还来可以根据虚假评论的比例判断平台的信用状况。

另一方面，不仅仅是平台和商家的层面。对于消费者来说，虚假评论的标注，能够帮助消费者清晰的识别虚假评论，客观的了解卖家的信用状况，更能够让在线评论发挥它应有的价值和作用，为消费者购物提供参考和意见，帮助消费者快速、准确地做出选择。

1.2 国内外文献综述

通过阅读国内外相关文献资料，发现对于虚假评论的识别主要从两个角度入手：

1.2.1 评论文本信息角度

第一是从评论文本信息角度对虚假评论进行识别，现有的研究当中主要是寻找合适的分类模型并设计特征来提高模型的分类效果。大致可以分为三种模式：半监督、无监督和有监督的学习方法。

半监督也是属于机器学习领域的某一类算法，无监督学习是指没有标注数据，模型根据数据特性进行学习，有监督是指数据是标注完好的，模型根据标注数据进行学习，而半监督介于这二者之间，将大部分没有标注的数据和少部分有标准的数据相结合，当标注数据较少的时候适合应用这种方法。Li(2011)等[2]人将半监督方法当中的协同训练方法应用在虚假评论的识别当中，并且使用大量的未标注数据集进行计算，结果表明基于半监督的协同训练算法能够很好的处理虚假评论识别问题。任亚峰（2014）[22]等针对已有的虚假评论方法的不足，设计出 2 种半监督算法对大量没有标注的文本充分利用，并整合了计算机语言学和心理语言学的特点，组合最好的特征组合，实验结果表明，具有很好的效果。杜茂康（2019）[15]将主成分分析与半监督的协同训练算法相结合，用以解决虚假评论识别数据难以获取的问题，并且从评论者以及评论行为两个角度出发，构建一套基于虚假评论识别的指标体系，并进行虚假评论的识别。杨云云（2020）等[12]人利用半监督的 tri-Training 算法对在线电影评论进行虚假评论的识别，并与支持向量机的分类结果进行对比，发现在实践过程中，半监督学习也能较好的完成虚假评论检测的任务。

正如前文所提及的，无监督使用无标注数据，对数据特征进行提取。2011 年，Raymond（2011）等[8]针对 Amazon 现实评论数据集，提出将语义模型和文本挖掘相结合，并将其结果并与 SVM 相比较，实验结果表明无监督的学习方法可以较好的识别虚假评论，但是该试验只是将有高度相似性的评论标记为虚假评论。顾松敏（2018）[20]提出一种 GSLDA 的方法来对虚假评论进行识别，这种方法基于传统的 LDA 方法，以三个阶段排查出虚假评论当中的作弊群组，并且证明了该方法的有效性。

有监督的学习方法需要利用大量标注数据进行模型训练，具有很好的分类效果。Nitin（2008）[3]等人对虚假评论进行识别，其根据数据集进行特征抽取，并采用逻辑回归的有监督的识别方法对虚假评论进行识别，并取得良好的效果，但

是该研究主要是基于英文评论文本进行相关的研究。C.L.La（2010）等[4]采用统计语言模型构造虚假评论的特征，该研究使用一元语法模型，并假设各个词汇之间相互独立，并没有考虑到上下文词语之间的隐含语义关系，研究的结果不是很理想。Myle Ott（2011）等[5]构造了一个黄金数据集，将其应用于虚假评论检测和识别。该研究结合计算机、心理学和语言学等理论创建出特征集合，并利用这些特征构造支持向量机分类器，对虚假评论进行识别，完成虚假评论识别任务，这些人为构造的特征并没有很好的反映出文本内部的语义关系。

集成学习方法能够提高模型的稳定性，汪浩（2020）等[13]提出了一种基于集成学习的虚假评论识别方法，并将其应用在 Yelp.com 数据集上，并与传统的机器学习模型相比较，实验发现集成学习具有更高的精度和稳定性。

近年来，随着神经网络算法的不断发展，学者们利用神经网络算法在语音、图像等领域取得了很大的成果，基于神经网络对自然语言的处理也取得了长足的进步，越来越多的专家、学者将深度神经网络的方法应用到自然语言处理当中。

张胜男（2016）[18]等通过神经网络的方法实现对虚假评论的识别，将虚假评论的识别看作是一个二分类任务，其针对亚马逊数据集，通过使用神经网络的方法对文本自动提取特征，并于传统的机器学习分类器 SVM 进行结合实现虚假评论的识别，实验结果与传统的特征提取方法进行对比，发现基于深度学习的方法能够更好的提取文本特征。Ren（2017）等[7]从词粒度出发，利用神经网络来对句子编码处理，构造句向量，并将其处理成为文本特征向量，进行虚假评论的识别，根据其实验结果可以看到，这种方法要优于当时的其他方法。李静（2017）等[21]对现有的处理文本分类的方法做出总结，利用主题-动态卷积神经网络算法解决一般的文本分类处理方法无法有效处理文本特征的问题，并通过这种方法提高了识别的准确率。韩俣谈（2018）等[19]通过在卷积神经网络当中使用不同大小的卷积核来避免文本信息的流失问题，实验的结果表明，这种方法能够很好的处理虚假评论识别的问题，在各项指标上表现都很优异。刘秀（2019）等[14]人利用双向 LSTM 和注意力机制对评论文本抽取特征，建立虚假评论识别模型，该实验的结果表明圣经网络方法能够很好的提取文本特征，在文本分类问题上取得不错的效果。

1.2.1 评论者角度

第二是从评论者或者刷单团体的角度对虚假评论进行识别，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。Lim（2010）[9]针对 Amazon 数据集，针对性地设计了几种行为特征，并对用户行为利用线性加权的方法进行评分，根据评分来检测出虚假评论者。Wang（2014）等[6]提出一种图模型，通过获取商家、消费者、评论内容这三点之间的关系，根据模型计算信任得分，根据信任得分来对虚假评论进行识别。张文宇（2018）[16]针对亚马逊中国电商网站的评论者数据对虚假评论者的行为动机进行分析，并构建 D-S 证据理论模型，并结合 SVM 对虚假评论进行检测，具有一定的准确性，但是用户的行为偏好是动态的，容易受到外界事物影响，所以具有不稳定性。曹文盼（2017）[21]基于特定品牌的虚假评论者识别办法，构建评论用户关系网，使用多特征尺度模型，并构建水军检测模型用于检测，并根据评分差异、评论产品目标差异等对水军的阵营进行分析，研究此方法对于水军阵营发现较为有效。

1.2.3 文献述评

综上所述，识别虚假评论主要有两种方法，从评论文本信息出发和从评论者行为特征出发。但是，不同的平台，虚假评论者的行为模式是不同的，我们所得到的评论数据也会有差别，具有不同的平台特征，所以使用这种方法得到的模型的效果相对来说比较差，而且关于评论者的数据相对来说难以收集，相比较而言，评论文本数据的获取要方便的多。所以，本文将从文本评论数据出发，对文本评论数据进行建模、识别虚假评论。

表 1.1 研究方法对比

	方法	识别效果	方法复杂度	说明
无监督	语义分析	低	中	误判率高
	聚类法	中	中	超参数无法确定
半监督	Pu 学习	高	高	必须满足数据某种假设
	Co-training	中	高	要求两视图的特征具有独立性
有监督	LR	中	低	容易欠拟合
	SVM	高	低	适用于处理高维特征问题
	RF	高	低	适用于处理不平衡数据问题，减少过拟合
	NB	高	低	类条件概率独立假设太严格
	集成学习	高	高	提升模型的稳定性，数据不平衡能够胜任
	神经网络	高	高	自动学习文本特征

基于评论的文本内容进行虚假评论识别主要由上述三种方法，我们可以通过对比看到，在这三种方法当中，有监督的识别方法是最常用的，并且稳定性比较高，所以本文主要采用有监督的识别方法对数据进行建模。在有监督的学习方法的当中，SVM 是研究者们最常用的方法，另外可以看到集成学习方法能够提升模型的稳定性，通过使用神经网络的方法建模不需要人工构建文本特征，它能够自动地学习文本特征，并进行反向传播学习。本文采用传统的机器学习方法和深度学习方法相结合的方式识别虚假评论，并且利用集成学习方法对单个基分类器进行组合，提升模型的泛化能力、稳定性。

目前虚假评论识别当中存在的一些问题：

一是之前对于虚假评论的检测研究大多数都是通过经验，人工地去构造文本特征，利用这些构造的特征去建模分析，预测的结果会受到构造的文本特征的影响。

二是之前对于虚假评论的研究主要是使用单个分类器对评论文本进行分类计算，模型的稳定性、泛化能力比较差。

三是现有的虚假评论的研究主要是对评论文本做虚假评论识别，没有考虑到评论语义、评论、评论者和商家之间的潜在关系，对虚假评论行为模式进行特征提取。

四是现有的虚假评论检测多用于英文评论文本的识别，中文领域的研究较少。

1.3 研究的主要内容，重点解决的问题，预期结果

1.3.1 研究的主要内容

本文针对目前虚假评论识别当中存在的一些问题，本文虚假评论的识别主要包括五个部分：

第一，首先获取数据。利用爬虫技术，获取某电商平台某品牌所有店铺的评论数据。

第二，构建数据集，对数据进行清洗。对爬取的数据进行数据清洗，去除无效评论。选取 5-8 名电子商务研究生，阅读现有文献分析虚假评论的特点，进行人工标注虚假评论，当人工出现分歧，采取少数服从多数的原则确定最终的标注结果。

第三，文本数据的预处理。进行数据清洗，随后利用现已经非常成熟的中文分词技术对文本语料进行分词处理，将一段语料切分成一个个单独的词汇，并利用 word2vec 工具构建词向量。

第四，构建虚假评论识别模型，本文采用传统的机器学模型和神经网络模型相结合的方法构建模型。传统的机器学习模型，如 SVM、朴素贝叶斯、逻辑回归等在文本建模当中的表现都比较好，而神经网络模型能够弥补机器学习模型需要人工构建文本特征的缺陷，具有自动学习文本特征的能力。所以本文次啊用两种相结合的方法对数据进行建模分析，最终再利用集成学习的方法对模型进行融合，提升模型的稳定性。

第五，虚假评论特征分析。在完成模型的搭建后，对所有的未标注数据进行分类，对两类评论进行特征分析，如词云展示、n-gram 模型高频词汇组合展示、两类文本评论的词性对比、LDA 主题差异分析、评论情感分析等。

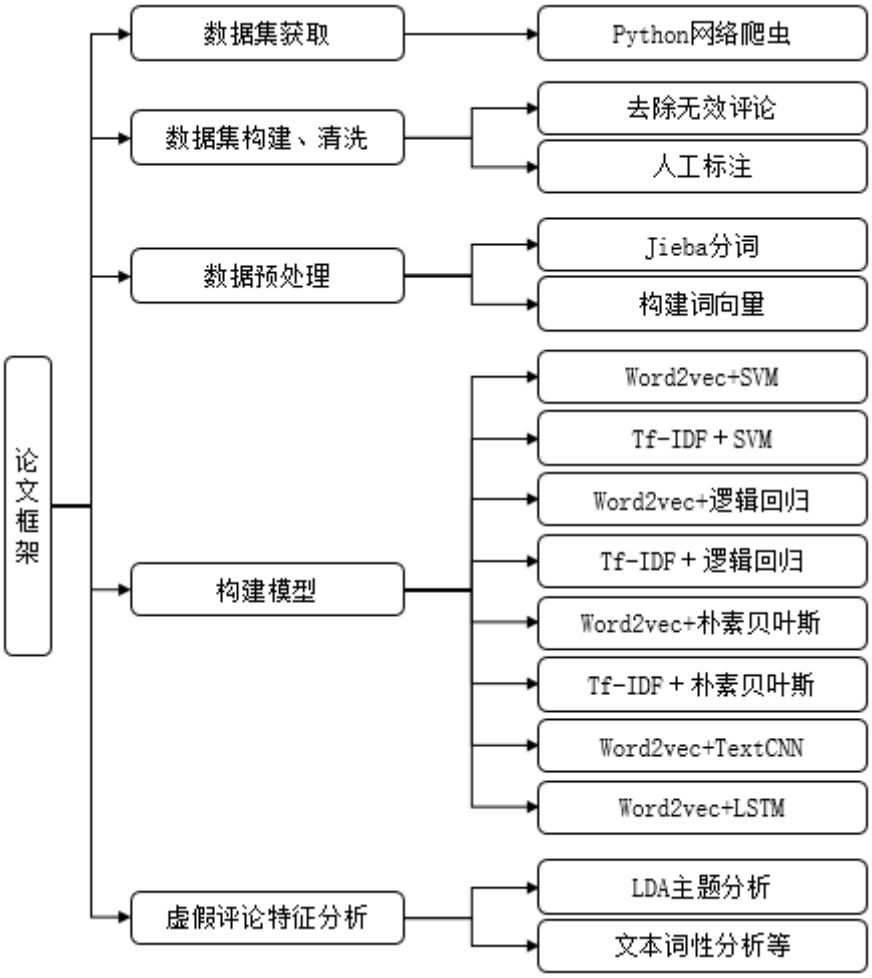


图 1.1 研究思路

1.3.2 重点解决的问题

一是对非结构化的文本数据集构建词向量，包括 TF-IDF 和利用 word2vec 的 skip-gram 进行词向量的构建；

二是利用神经网络模型 LSTM 和 Text－CNN 对评论文本特征进行抽取，结合不同分类器对评论文本进行分类，选取分类效果最好的分类器组合，并进行模型融合，对数据进行预测；

三是对分类文本进行特征分析，对比真实评论和虚假评论的各项特征，对虚假评论行为模式进行分析。

1.3.3 预期的结果

（1）将各个训练好的基分类器利用集成学习的方法进行模型融合，准确抽取各类的文本特征，弥补单个模型的不足，提升模型的稳定性，实现大规模、准确地预测评论的类别。

（2）对评论文本进行特征提取，对比真实评论，通过对比 lda 主题差异等方法，分析出虚假评论的语言特征以及行为特征和真实评论之间的区别，总结出虚假评论所具有的特性，可以作为以后相关研究提供参考。

1.4 可能的创新点

本文的第一点创新在于模型。之前的研究，学者们多以单个机器学习分类器进行虚假评论的识别，本文采用了传统机器学习分类模型和深度神经网络模型相结合，并且采用两种特征向量化的方法，以此构建最终的虚假评论识别模型，不仅将不同的特征向量化方法于不同的模型相组合，选择出最优的模型，并利用集成学习的方法进行模型融合，提升模型的稳定性。

本文的第二点创新是不仅构建了虚假评论识别模型，还通过 lda 和主题差异对比、语言模型、情感分析、词性分析等方法，并结合评论者其他维度的信息，考察虚假评论的语言特征和行为特征。

2. 文本挖掘相关技术及理论

文章的此部分将会对虚假评论识别过程中所涉及到的一些相关技术和理论进行阐述和介绍，主要介绍数据挖掘的相关理论和相关的算法。

对于虚假评论的文本挖掘一直以来都是虚假评论检测的热点研究内容。评论文本作为一种非结构化的数据，包含着非常丰富的信息和内容，表达用户和消费者对于商品信息的态度和使用感受，可以通过文本挖掘的方法发现其隐藏的信息。

文本挖掘是一个多学科交叉的领域，多个领域的知识，本章主要介绍文章后续需要的重要技术和相关概念。

2.1 文本预处理

本文所研究的主题是评论文本数据，它属于一种非结构化的数据，为了让计算机理解这种数据，需要对文本数据进行预处理，让它成为一种计算机可以理解的形式，变成一个个词汇，中文文本数据的预处理一般分为以下几个部分：

去除非文本的部分：过滤掉文本中存在的非法字符和对文本语义分析产生干扰的符号(包括表情符号、标点符号、数字和网络链接等),此部分可以通过python的re库正则表达式来实现操作。

去除停用词：文本当中包含着诸如的、了、吗等频率较高但对文本语义不会产生影响的词汇，需要对其进行处理。

中文分词：中文的语法、语言习惯以及文本形式与英文有着很大的不同，在英文的文本当中，每个词汇都以空格间隔开来，有空格作为切分点，但是中文不可以，所以需要进行分词，将一段文字切分成为一个个词的形式。近年来，中文分词技术已经非常成熟，有很多工具可以实现文本分词，本文主要采用 jieba 分词工具的精确模式对文本内容分词。

2.2 文本的表示方法

在文本的表达当中的经典方法是空间向量模型，在普通的文本分析当中也会被称作词袋模型。词袋模型作为文本表示的常用模型，不考虑句子的先后顺序问题，当然也不会考虑到句子的现实语义，通常用于长文本分析或者信息检索当中，

但是对于短文本分析来说，主要依靠句子当中比较关键的几个词语，所以词袋模型并不适合这种场景，它并不能满足这种需求，与此同时，文本的分布式表示方法提供了一种比较适合的解决方法。

正如之前所提到的，词袋模型并不会注意到词语之间的顺序和上下文语意，但是分布式表示方法就解决了这个问题。词语的分布式假说由 Mikolov 等人基于神经网络模型提出的一种能够有效训练词向量的模型。

词语的分布式假说是指：一个词语的语义由其上下文决定，即上下文相似的词语，其语义也十分相似。近年来，随着数据规模的不断扩大、相关技术的发展，分布式表示方法越来越受到学术界以及工业界的重视，被广泛应用到研究和现实场景当中，分布式的表示方法的优点在于，它可以使用一组定长的向量来表示某个词汇，如果某个词汇的含义与之相近，那么它们向量之间的距离也会比较近。

Word2vec 是一种词的分布式表示应用的词向量工具，它可以实现将某个词汇表示成一个具体的向量，并且利用向量之间的空间距离大小表示词汇之间语义的相近程度。Word2vec 可以通过两种方式实现，分别是 skip-gram 和 CBOW。

首先介绍 CBOW 模型，它的基本思想是：输入上下文词语开预测中心目标词语，以目标词的概率为优化目标。

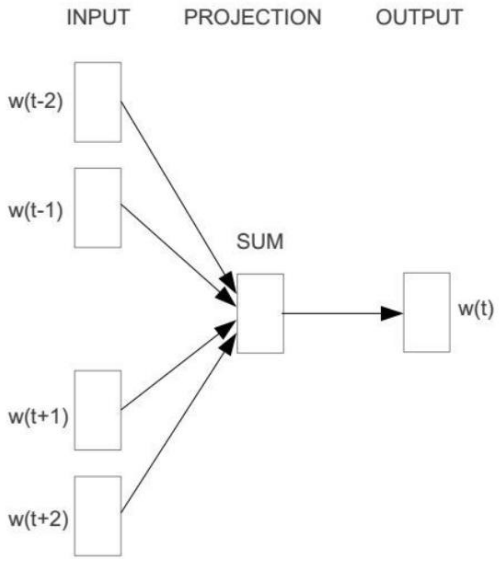


图 2.1 cbow 模型图

CBOW 模型的计算如图 2.1 所示，结构主要分为输入层、映射层、输出层，并且采用 logistic 回归的形式计算中心目标词的概率。

与 CBOW 利用上下文来预测中心词的方法刚好相反，Skip-gram 采用中心词

来预测上下文的词，图 2.2 展现 Skip-gram 的基本思想。

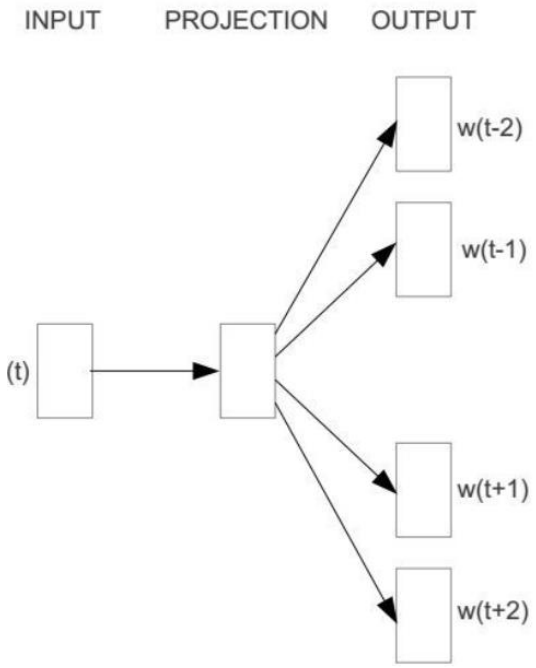


图 2.2 skip-gram 模型图

Skip-gram 的目标函数如下所示：

$$L^* = \underset{L}{argmax} \sum_{w_i \in V} \sum_{w_c \in Wc} logP(w_c|w_i) \tag{2.1}$$

上式中，V 表示词表的大小，w_i表示中心词，w_c表示窗口为 c 的上下文背景词。我们的目标就是最大化这个似然函数，找到合适的 L 矩阵，来表示词向量矩阵。

不论是 CBOW 还是 Skip-gram，它们计算的复杂度与词表的大小相关，词表越大，那么计算复杂度越高。为了提高计算效率，一般会采用负采样和层序 softmax 方法来提高计算效率。

2.3 LDA 主题模型

LDA 主题模型即潜在狄利克雷分布（LDA）主题模型,由David等提出，它属于无监督学习的一种。利用 LDA 主题模型，可以通过聚类的形式，挖掘出文章或者数据隐藏的主题信息，具体算法过程如图 2.3 所示：

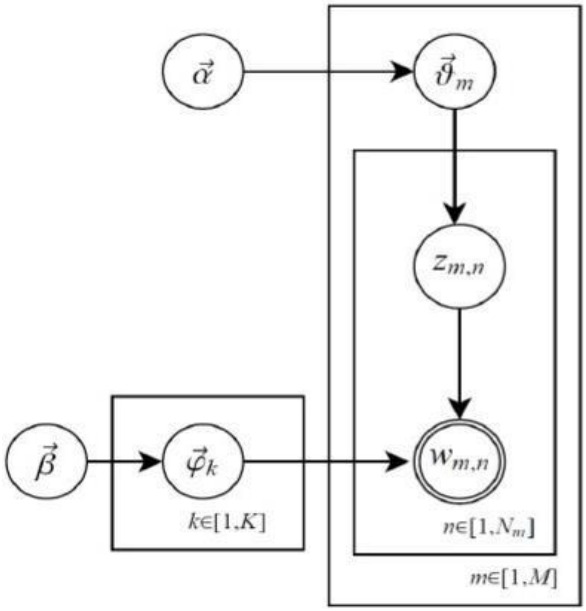


图 2.3 LDA 主题模型图

如图所示，箭头表示两变量间地条件依赖性，方框表示重复抽样，方框右下角表示抽样地次数，具体生成过程如下所述：

（1）从狄利克雷分布α中取样生成文档的m主题分布θ。

（2）从主题的多项式分布θ中取样生成文档m第n个词的主题Z。

（3）从狄利克雷分布β中取样生成主题Z的词语分布φ_k。

（4）从词语的多项式分布φ_k中采样最终生成词语w。

表 2.1 LDA 模型的主要参数

符号	含义
M	文档个数
K	主题个数
V	词项个数
α	θ_m的先验分布超参数（K 维向量）
β	φ_k的先验分布超参数（V 维向量）
θ_m	第 m 个文档的主题分布参数
φ_k	第 k 个主题的词项分布参数
N_m	第 m 个文档的长度
$z_{m,n}$	第 m 个文档第 n 个词对应的主题
$w_{m,n}$	第 m 个文档第 n 个词对应的词项
$z_m = \{z_{m,n}\}_{n=1}^{N_m}$	第 m 个文档对应的主题序列
$w_m = \{w_{m,n}\}_{n=1}^{N_m}$	第 m 个文档对应的词项序列
$w = \{w_m\}_{m=1}^{M}$	文档集对应的词项序列
$z = \{z_m\}_{m=1}^{M}$	文档集对应的主题序列

2.4 支持向量机

支持向量机（support vector machines，SVM）是一种非常经典的机器学习分类器，于90年代中期Vapnik和他的At&Bell实验小组提出，根据样本分布情况的不同，可以分为三种分类器：硬间隔线性可分支持向量机、软间隔线性支持向量机和非线性支持向量机，根据样本分布的不同选择合适的分类器，另外，因为使用凸二次规划，一般能够找到全局最优解。

下面简单介绍线性支持向量机。

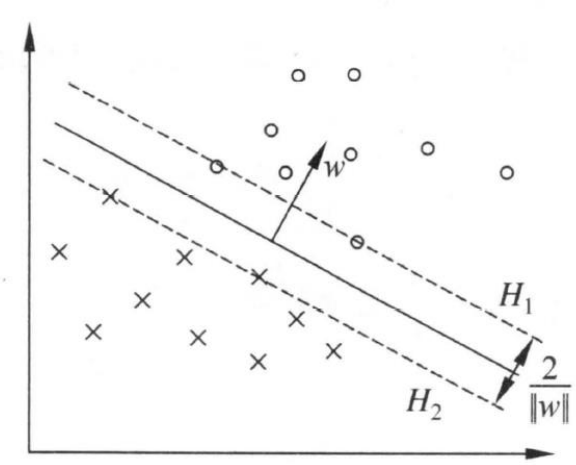


图2.4 线性可分

如图2.4所示，如果存在一个超平面将样本完全分开到超平面的两侧，那么就称之为样本完全线性可分，在这种情况下SVM通常被描述称为一个带约束的优化问题。

$$\min_{w,b} \frac{1}{2}\|w\|^2 \tag{2.2}$$

$$s.t. \quad y_i(w \cdot x_i + b) - 1 \geq 0, \qquad i = 0,1,2,\ldots,N \tag{2.3}$$

对于这种凸二次规划的问题，可以利用拉格朗日乘子法并结合对偶性，得到参数w，b的最优解w^*和b^*，由此得到最优的分离超平面：

$$w^* * x + b^* = 0 \tag{2.4}$$

而且得到的参数是目标函数的最优解。

得到分类决策函数：

$$f(x) = sign(w^* * x + b^*) \tag{2.5}$$

对于样本近似线性可分的情况，采用最大化软间隔的方式，对训练数据进行

训练。

$$\min_{w,b}\frac{1}{2}\|w\|^2+C\sum_{i=1}^{N}\xi_i \tag{2.6}$$

$$s.t.\quad y_i(w\cdot x_i+b)-1+\xi_i\geq 0,\qquad i=0,1,2,\dots,N \tag{2.7}$$

$$\xi_i\geq 0,\qquad i=1,2,\dots,N \tag{2.9}$$

得到最优参数及分类决策函数：

$$f(x)=sign(w^*\ast x+b^*) \tag{2.10}$$

对于样本数据完全线性不可分的情况，采用核技巧的方式，将原始空间的数据映射到一个新的特征空间，在这个新的特征空间当中学习分类模型。

$$K(x,z)=\ \phi(x)\cdot\phi(z) \tag{2.11}$$

称$K(x,z)$为核函数，$\phi(x)$为映射函数，式中$\phi(x)\cdot\phi(z)$为$\phi(x)$和$\phi(z)$的内积。选择适当的核函数$K(x,z)$，求解得到的就是非线性支持向量机

$$f(x)=sign\left(\sum_{i=1}^{N}{\alpha_i}^*y_iK(x,x_i)+b^*\right) \tag{2.12}$$

2.5 Logistic Regression

逻辑回归（logistic Regression）是一种统计学习当中的经典方法，作为一种典型的二分类模型被人们所熟知，由概率分布$P(Y|X)$表示。其中，随机变量 X 的取值为实数集，而随机变量 Y 的取值为 0 或者 1。二项逻辑斯蒂克模型是如下的条件概率分布：

$$P(Y=1|x)=\frac{\exp(w\cdot x+b)}{1+\exp(w\cdot x+b)} \tag{2.13}$$

$$P(Y=0|x)=\frac{1}{1+\exp(w\cdot x+b)} \tag{2.14}$$

在这里x是输入，$Y\in\{0，1\}$是输出，w和b是参数。对于一个固定的输入x，通过上面两个公式的计算可以得出 Y 属于 0 或者 1 的概率各自有多大，再继续比较这两个概率的大小，得出输入x属于哪个类别。

在模型的训练过程中，对于一组给定的数据集$\mathrm{T}=\{(x_1,y_1),\dots,(x_n,y_n)\}$可以使用极大似然估计法来估计模型中的参数，以此得出逻辑斯蒂克回归模型。

假设：

$$P(Y=1|x)=\pi(x),\qquad P(Y=0|x)=1-\pi(x) \tag{2.15}$$

其似然函数为

$$\prod_{i=1}^{N}[\pi(x_i)]^{y_i}[1-\pi(x_i)]^{1-y_i} \tag{2.16}$$

其对数似然函数为

$$L(w)=\sum_{i=1}^{N}[y_i\log\pi(x_i)+(1-y_i)\log(1-\pi(x_i))]$$

$$=\sum_{i=1}^{N}[y_i(w\cdot x_i)-\log(1+exp(w\cdot x_i)] \tag{2.17}$$

对其求最大值，便能得到w的估计值。

2.6 朴素贝叶斯

朴素贝叶斯是一种生成模型，由英国数学家托马斯•贝叶斯提出，是一种基于贝叶斯定理和类条件独立性假设的分类算法。朴素贝叶斯方法对条件概率分布作了类条件概率独立性的假设，具体地，类条件独立性假设是指

$$P(X=x|Y=c_k)=P(X^{(1)}=x^{(1)},\dots,X^{(n)}=x^{(n)}|Y=c_k)$$

$$=\prod_{j=1}^{n}P(X^{(j)}=x^{(j)}|Y=c_k) \tag{2.18}$$

后验概率：

$$P(Y=c_k|X=x)=\frac{P(Y=c_k)P(X=x|Y=c_k)}{\sum_k P(Y=c_k)P(X=x|Y=c_k)}$$

$$=\frac{P(Y=c_k)\prod_j P\left(X^{(j)}=x^{(j)}\middle|Y=c_k\right)}{\sum_k P(Y=c_k)\prod_j P\left(X^{(j)}=x^{(j)}\middle|Y=c_k\right)},\qquad k=1,2,\dots,K \tag{2.19}$$

根据朴素贝叶斯分类的基本公式，朴素贝叶斯的分类器可以表示成如下的形式：

$$y=f(x)=arg\max_{c_k}\frac{P(Y=c_k)\prod_j P\left(X^{(j)}=x^{(j)}\middle|Y=c_k\right)}{\sum_k P(Y=c_k)\prod_j P\left(X^{(j)}=x^{(j)}\middle|Y=c_k\right)},\qquad k=1,2,\dots,K$$

$$=arg\max_{c_k}P(Y=c_k)\prod_j P\left(X^{(j)}=x^{(j)}\middle|Y=c_k\right) \tag{2.20}$$

根据后验概率的大小，判断x属于哪个类。

2.7Text-CNN

Text-CNN模型是利用卷积神经网络对文本进行分类的算法,2014年,ToonKim再其论文中提出利用CNN对文本进行分类，其结构如下图所示。

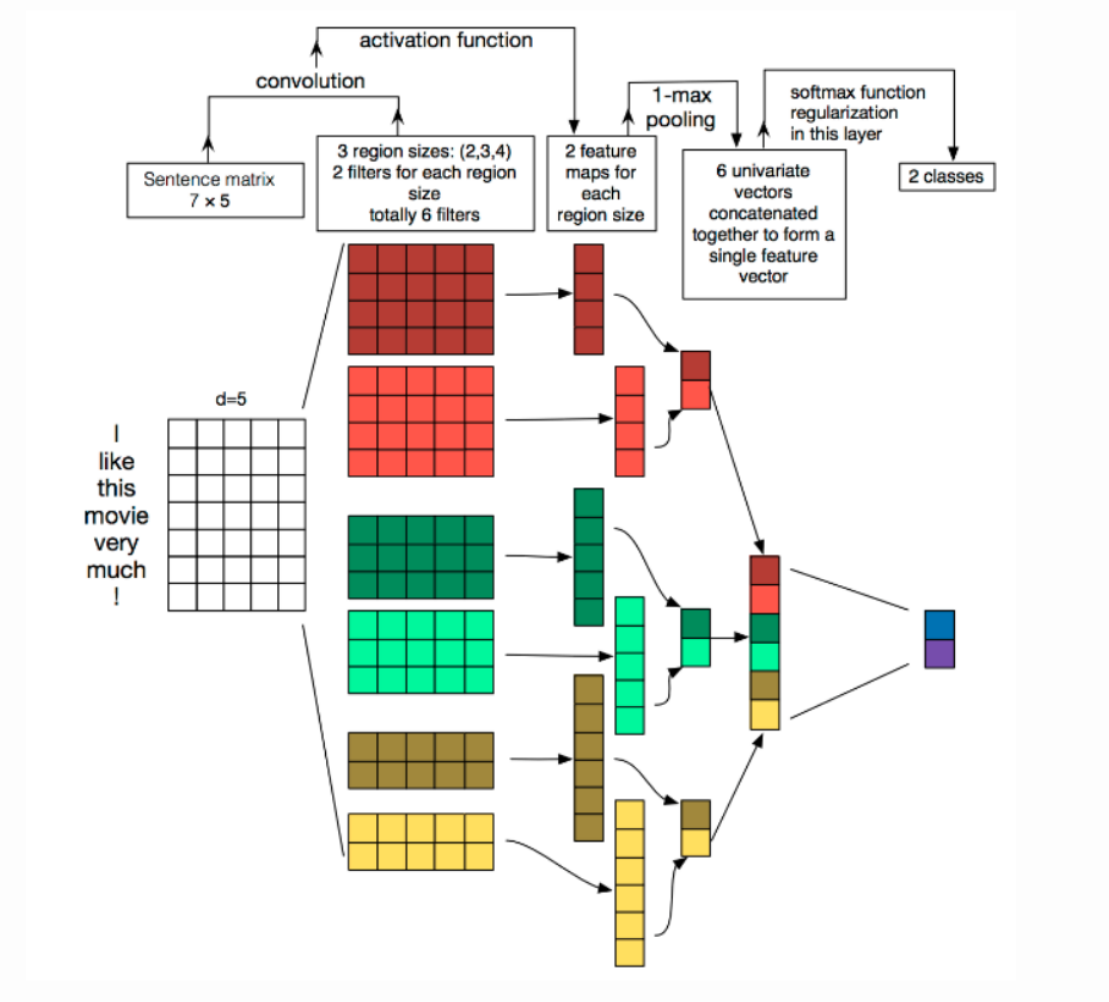


图 2.5 Text-CNN

它主要分为嵌入层、卷积层、池化层以及输出层。

首先是嵌入层，也叫做 embedding layer，这一层一般是将已经预训练好的向量放进去，会形成一个矩阵 A，这个矩阵 A 的每一行都是一个词向量，另外，这个矩阵 A 可以是静态的也可以是非静态的，如果是静态的，那就会保持固定不变,但是如果是动态的,那么就可以对这个矩阵进行反向传播,更新里面的参数。

随后是卷积层，如果我们的目标是一个完整的句子，那么首先会对句子进行分词的操作，假设这个句子分词之后有 k 个词汇，词向量为 d 维，那句子里面的每个词语与嵌入矩阵 A 可以得到一个词向量,对于这个句子可以得到 k 行 d 列的

矩阵。对于这个矩阵将其看作是一张图像进行卷积处理，提取特征，因为句子当中的词语上下文关联性特别高，所以采用一维卷积处理，并不像普通的图像卷积处理方法，文字卷积旨在文本序列上进行，词向量的维度是不变的，学习文本的不同特征，由此可以得到特征图。

池化层：由以上步骤，不同大小的卷积核所得到的特征图的大小是不同的，要对这些特征图的维度进行统一，对每个得到的特征图进行池化操作，池化函数一般有最大池化和平均池化。对于最大池化来说，最常用的就是 1-max pooling，提取出特征图当中的最大值，来选择最重要的特征，对所有的卷积核，做 1-max pooling，在联合起来，就可以得到特征向量，将这个向量输入到 softmax 层进行分类。对于平均池化的方法来说，顾名思义就是取特征图的均值。

2.8 长短期记忆网络

首先介绍循环神经网络（RNN），1982 年，美国加州理工学院物理学家 John hopfield 发明了一种单层反馈神经网络 Hopfield network，用来解决组合优化问题，这是最早的 RNN 的雏形。传统的神经网络并不能利用序列之前的小信息，随着信息的传递，以往的信息会被丢失，但是循环神经网络可以通过对某个信息进行循环来达到储存信息的目的。

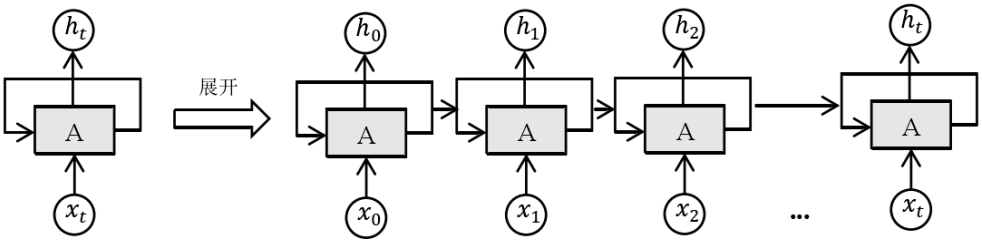


图 2.6 循环神经网络

在上图中，可以看到，可以将 A 看作是一个神经网络，他的任务就是接收 x_t，并输出 h_t，可以看到存在一个循环结构实现了信息的循环和利用。图的右边部分，将这个网络结构进行展开，可以清楚的看到信息的传递，将当前网络的输出传递给下一层。但是传统的 RNN 遇到长序列的问题的时候，需要保留很久之前的信息的时候，就表现的捉襟见肘了，RNN 难以解决这种长依赖的问题。

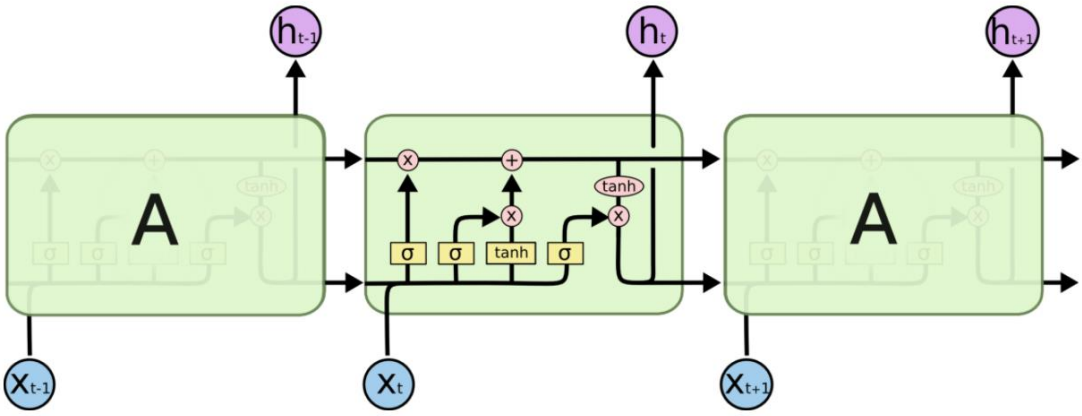


图 2.7 LSTM 模型

长短期记忆网络（LSTM）是传统的 RNN 的一种改进方法，1997 年，Jurgen Schmidhuber 提出，该网络的设计的初衷就在于解决长依赖的问题，被广泛应用。LSTM 的核心就在于它存在一种细胞状态，它就像一个传送带一样，存在于整个网络的始终，将历史信息进行传递。当然，并不是所有的信息都会被保留下来，LSTM 有一种“门”结构，对这个细胞状态进行添加或者删除信息的操作，这种“门”结构一般是由点成操作和激活函数组成，来对信息操作。

它有三个门，分别为输入门、遗忘门以及输出门，各司其职。首先，神经网络需要决定丢弃哪些信息，这个部分主要是通过“遗忘门”，进行操作，通过一个 sigmoid 激活函数，来决定保留或者丢弃多少。随后是决定为细胞状态添加哪些新的信息，这一步需要候选细胞状态进行配合。然后更新旧的细胞状态，产生新的细胞状态，通过遗忘门和输入们协同进行更新。更新完细胞状态之后，需要判断输出哪些状态，通过输出门的 sigmoid 激活函数进行操作。

2.9 集成学习方法

集成学习方法是一大类统计机器学习方法的统称，此方法当中包含了很多集成学习的思想，其中最主要的是 Bagging 和 Boosting 这两分类方法。

Boosting 在训练模型的时候，采用串行的方式训练学习器，将各个基分类器层层叠加，它们之间有着强依赖存在。在每一层的分类器训练的时候，给予前一层分类错误的样本更加高的权重，在最终的预测当中，将各层分类器的结果加权

来得到最终分类器的输出结果，在 Boosting 方法当中，最具代表性的就是梯度提升决策树（BGDT）。

Bagging 方法和 Boosting 方法不同在于各个分类器在训练的时候可以采用并行的,各层分类器之间没有强依赖性。在 Bagging 当中最著名的莫过于随机森林（Random Forest），它为了让各个分类器之间相互独立，将训练数据集分成为若干个子集，它更像是一个集体决策的过程，每个分类器单独做出决策，到最后各个分类器投票表决，得到最终的结果。

3. 数据的采集与预处理

3.1 数据的采集及多维特征分析

本文研究的目的在于探索出一套完整的、精确度高的、具有可推广性的虚假评论的识别办法。文本数据来源自京东商城某品牌手机商品评论，在京东商城搜索某品牌手机，选择了 6 家店铺的评论数据，并完成了相关数据的采集，共收集到原始数据 44679 条，并对评论数据进行建模分析。

京东商城的评论数据具有以下的维度特征：会员名称、会员级别、评价星级、评价内容等。为了获取数据的方便、快捷，本文所使用的评论数据均是利用 python 设计网络爬虫获取而来，评论数据真实可靠。

会员	级别	评价星级	评价内容	时间	点赞数	评论数	追评时间	追评内容	采集时间
cgd182	PLUS会员	star5	外观很惊	2020-03-	188	92			2020-12
M***g		star5	首先，手	2020-09-	10	11			2020-12
俱***酌	PLUS会员	star5	帮人拿的，	2020-06-	33	9			2020-12
小小陈cw	PLUS会员	star5	买给婆婆	2020-05-	77	19			2020-12
h***2		star5	10天之内，	2020-02-	125	36			2020-12
jd_13734	PLUS会员	star5	宝贝已经	2019-12-	67	16	[购买6天后追评]		2020-12
jd_千里蚓	PLUS会员	star5	外形外观：	2020-03-	49	7			2020-12
j***1		star5	屏幕音效：	2020-04-	26	3			2020-12
**	PLUS会员	star5	外形外观：	2020-04-	25	3			2020-12
K***g		star5	Redmi红米	2020-06-	10	2			2020-12
吖-谁呀	PLUS会员	star5	买给丈母	2019-10-	159	54			2020-12
微***啊	PLUS会员	star5	手机外形	2020-06-	11	5			2020-12
刘志品za	PLUS会员	star5	红米8A给	2020-03-	39	13			2020-12
拼***官	PLUS会员	star5	629，还能	2020-06-	13	11			2020-12

图 3.1 原始数据部分样本

表 3.1 会员比例

index	级别	占比
PLUS 会员	21,669	0.48
PLUS 会员[试用]	269	0.01
非会员	22,741	0.51

对评论文本的会员维度进行统计，有 48%的消费者为 plus 会员，51%的消费者为非会员，在表 3.1 中我们可以看到，非会员用户和低级会员用户规模占到了

一般以上，用户质量中低等的占大多数，用户评论质量可能会出现参差不齐的状况。

表 3.2 用户评分

评价星级	数量
star5	44184
star1	225
star3	113
star4	98
star2	59

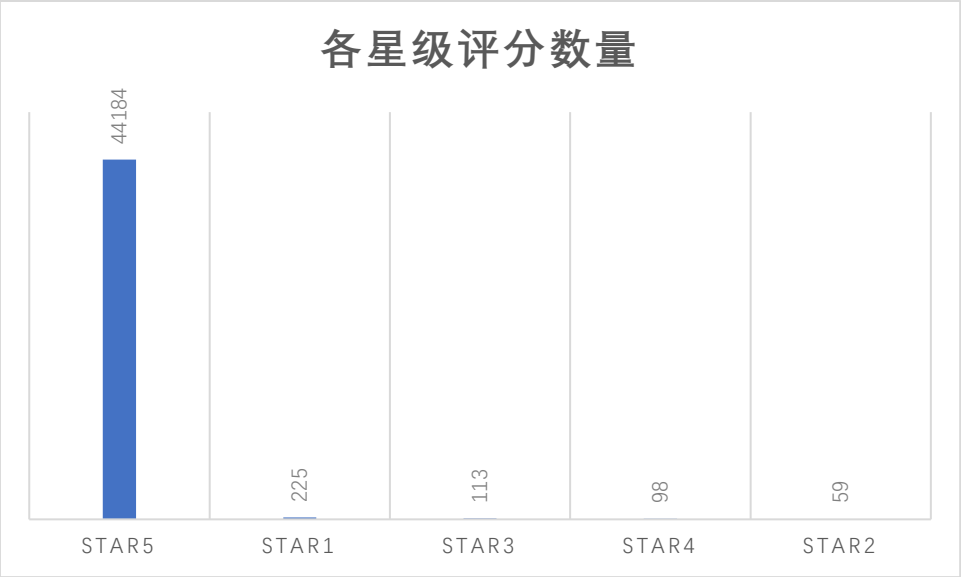


图 3.2 各星级评分数量

图 3.2 为京东评论用户评分各星级分布状况，有 44184 条评论的平均分为 5 颗星，可以看到绝大部分的用户都给了 5 星好评。

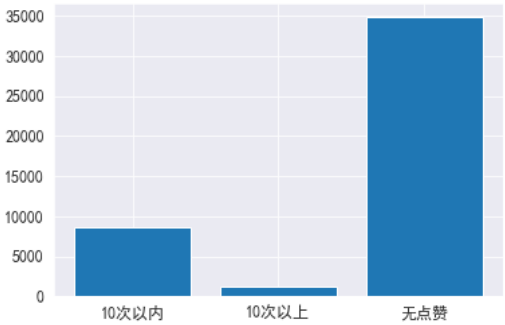


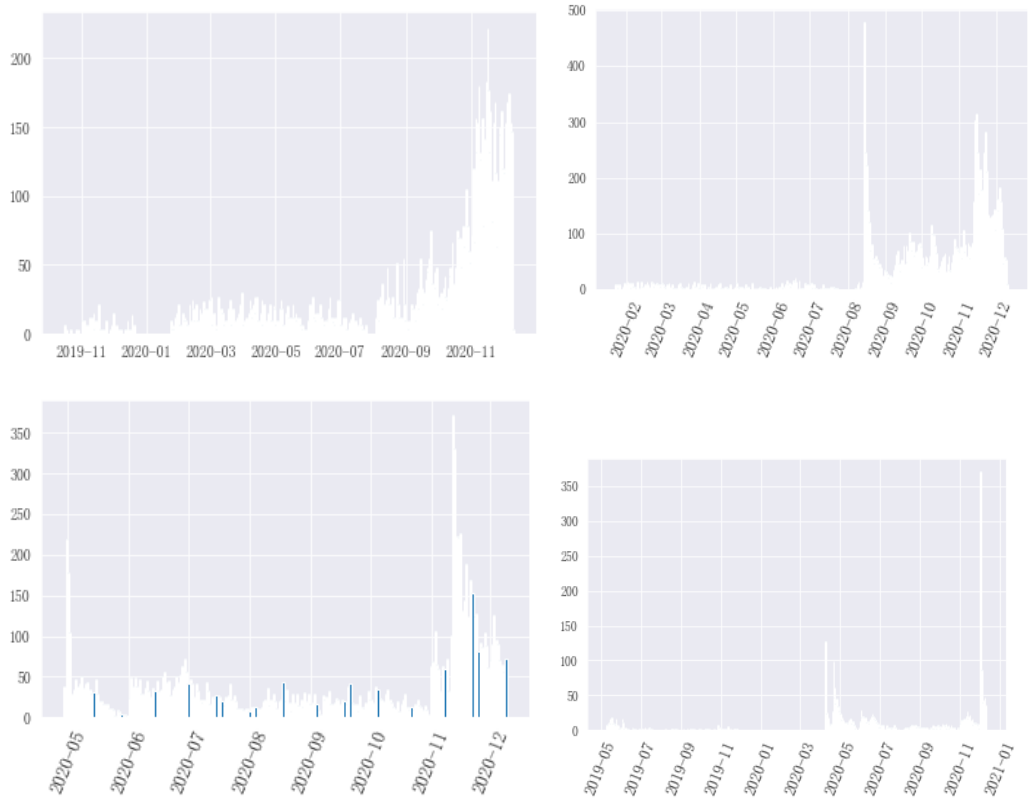
图 3.3 点赞次数

表 3.3 点赞次数

	counts	比例
10 次以内	8688	0.19
10 次以上	1195	0.03
无点赞	34796	0.78

对于评论文本被点赞次数进行统计可以看到，75%的评论数据没有被点赞，仅有 19%的评论被点赞 10 次以内，对于大部分评论数据，消费者可能觉得没有相应的参考价值。

对于某个电商平台来说，再排除掉电商节等降价活动之外，每个店铺的销量以及评论的数量应当稳定在某个值附近波动，而不会出现某种突然的巨大增加，如果该店铺的日评论数量长期稳定在一个水平，突然出现某日的评论数量激增，那么这一天很可能就会出现雇佣水军进行虚假评论等刷单行为，所以需要对数据集的每个店铺进行日评论数量统计，看是否出现异常行为。



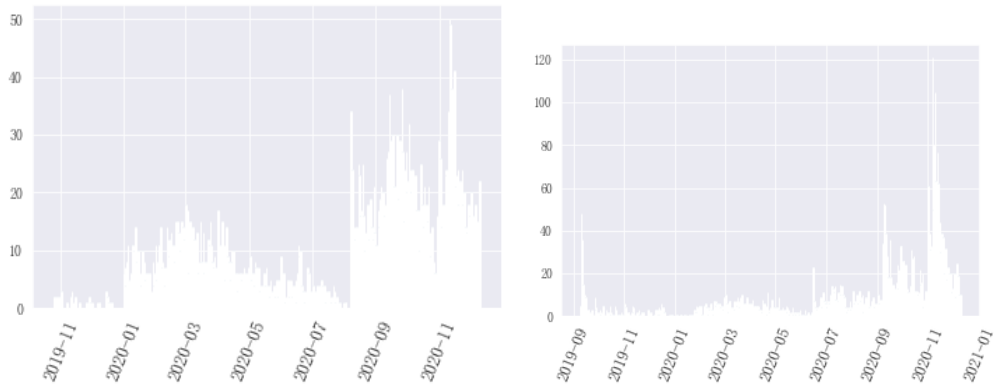


图 3.4 店铺评论数量分布

3.4 图为京东的 6 家店铺评论数量的时间分布图，每张图的横轴为时间，纵轴为评论数量。

表 3.4 各店铺评论数量描述统计

	店铺 1	店铺 2	店铺 3	店铺 4	店铺 5	店铺 6
count	422	323	225	582	412	458
mean	25.71801	37.41796	42.8	6.872852	9.662621	8.733624
std	40.10673	61.11959	50.04507	21.59909	8.94619	13.52083
min	0	0	2	0	0	0
25%	3	3	16	0	3	2
50%	12	9	28	2	7	4
75%	24	53.5	46	7	15	10
max	222	479	372	371	50	121
异常值阈值	105.9315	159.6571	142.8901	50.07104	27.555	35.77529

上表为京东 6 家店铺的每日评论数量的描述性统计。将高于评论数量均值两倍标准差的值定义为异常值阈值，如果日评论数量高于这个值，那么本文认为该日很可能会存在异常情况，需要进行标记。

通过上表的统计数据可以看到，各个店铺当中，均存在着高于异常值阈值的时间，对于这些日期内的评论很可能会出现虚假评论，所以需要对其标记，后续会结合评论的其他属性来具体判断其是否为虚假评论以进行数据标注。

表 3.5 异常评论店铺分布

店铺	店铺 1	店铺 2	店铺 3	店铺 4	店铺 5	店铺 6
异常评论天数	33	17	11	8	17	19

通过数据的对比，再根据每个店铺的异常值阈值，本文筛选出每个店铺高于异常值阈值的时间，如上表所示。这里的标记行为可以为后续的数据标注行为提供帮助，作为重要的标注依据。

3.2 数据预处理

本节主要介绍在社交网络平台当中对数据如何进行清洗以及文本句向量如

何构造。在做模型构建或者数据分析之前，对于数据的预处理和一写必要的数据清洗是很有必要的，在完成数据清洗之后，将每条数据转换成为计算机能够处理并且理解的数值形式，进行数据建模分析。

正如前面所介绍的那样，文本语言作为一种记录人们生活、感受的一种表现形式存在于内阁生活场景当中，它的主要用途就在于记录，人们通常会使用一串文字去记录自己的感受，随着手机和互联网的普及，这种类型的文本数据存在于QQ、微信、手机短信、微博及一系列的网络购物平台等当中，充斥者形形色色的文本信息。

本文所要研究的主体就是电商平台的在线评论，在这些评论内容当中，一定会存在着一些没有任何意义的文字和信息，需要对其进行剔除，避免其对建模结果产生大的影响，所以需要进行数据清洗。

一般而言，这类文本信息主要存在以下几个特点：

（1）文本字符长度比较短。根据相关的数据统计，人们在进行网络购物评价的时候，一般的文本长度大概在 100 字符左右，相比较一些文章、博客而言，这属于字符长度比较短的文本了，消费者倾向于使用比较短字符的文本评价来表达自己对于产品的使用感受。

（2）评论文本比较倾向于口语化。消费者在进行商品评价的时候，往往会和平时聊天一样，会添加一些 emoji 表情符号来表达自己的感受。

（3）文本存在非中文符号。消费者在进行商品评价的时候会不经意的使用一些网络用语来表达自己的观点。

基于以上几点，我们可以看到，为评论文本的局限性在于文本过短，这个特点会造成评论文本存在大量的影响实验结果的噪声，影响到文本向量化的结果表达，相应的也会影响到最终的模型结果。综上所述，评论文本的数据清洗十分必要。

对评论文本的数据预处理流程大致可以分为以下几步进行操作，分节进行阐述。

3.2.1 数据清洗

首先，商品的评论当中包含着很多的无效评价，比如系统默认好评等，这种

评论对后续的分析没有任何作用，并且可能会造成一定的噪声，所以本文针对此种无效评论一律剔除，并不会参与后续的分析过程。

在剔除掉前文所提出的无效评论之后，还需要针对评论内容进行进一步的处理，因为评论内容当中可能会包含一些难以辨别的符号和表情等，这种类似的评价对模型的构建会造成不好的影响，也需要进行剔除。本文基于以下规则对评论文本进行进一步的处理。

（1）文本完全由数字组成；

（2）文本完全由符号组成；

（3）文本完全由数字和符号组成。

若是文本评论满足上述三个条件，可以直接进行剔除。如果文本满足 str.isdigit()，则文本完全由数字组成；如果文本满足 str.alpha()，则文本全部都由字母组成；若文本满足 str.isalnum()，则说明文本完全由英文字母或者数字组成。本数据集有 3 条满足上述情况，将其剔除，并剔除 9 条默认好评，原始数据剩下 44667 条数据

3.2.2 文本去重

文本去重，字面意思很好了解，就是对评论文本进行去重处理。一般的网络购物平台评论会规定，如果消费者没有子啊一定的时间范围内对其所购买的商品做出相应的评价行为，那么平台就会自动对这些评论设置成为默认好评，然而这些评论并没有现实意义。

通过查阅相关的文献研究可以发现，研究者们对于文本去重有着很多深刻的研究，大量的研究者、专家使用比较复杂的算法去实现文本的去重，但是都会或多或少的存在着一些缺陷。比如，当我们使用编辑距离法计算两条语料文本之间的距离，编辑距离法首先会计算两条文本之间插入、修改、删除等编辑距离，然后认为设定一个阈值，如果两个语句之间的编辑距离小于设定阈值，则判断两个文本之间存在重复性。但是这种方法存在一定的局限性，比如“手机已经收到货了，试用了下，是运行速度很快”和 “手机已经收到货了，试用了下，是运行速度很慢”，二者的编辑距离为1，但是二者之间所要表达的意思是截然相反的，如果直接采取去重的操作，那么很可能会造成很多信息的流失，所以不能采取这

种策略。于此同时，在评论文本当中，消费者为了用尽可能少的字符表达自己的使用感受，这种类型的文本不在少数。所以本文摒弃了比较复杂的文本去重的方法，使用比较简单、直接的文本重复评论数据删除方法，针对重复的文本，只保留一条，以确保保留更多的原始语料信息。

表 3.6 重复评论

index	评价内容	时间
1448	蛮好的	2020/11/20 13:37
1459	经验	2020/11/1 10:11
1501	使用中	2020/11/25 19:03
1502	加价了	2020/11/14 22:11
1503	棒棒的	2020/11/6 11:47
1504	很棒	2020/12/7 10:59
1505	满意！	2020/11/28 20:34
1506	很好。	2020/11/12 8:02
1507	能行	2020/11/8 16:10
1508	ok 啦	2020/11/19 16:12
1569	太强了	2020/11/18 17:55

本文评论数据共有 44667 条，其中有重复数据 385 条。短文本出现重复的概率比较高，长文本不容易出现完全重复。如果在评论文本当中出现长文本，并且其重复次数比较多的情况，那么有可能是商家为刷手提供的一些固定的好评模板，当然，也有可能是消费者自己不想动手填写评论，而直接粘贴复制而形成，需要将其标记为虚假评论，但是在分析中这一类评论仍然需要进行去重处理。剔除重复的数据还剩下 44615 条数据。

3.2.3 短句删除

虽然比较精炼的辞藻能够比较方便、快捷地表达消费者的使用感受，但是在另一方面，比较简短的词汇难以覆盖消费者的消费感受，字数越少，所能够表达的语义信息也就越少。所以，要想表达一定程度上的使用感受，使用尽量少的词汇来做并不是一个很好的选择，过少的文本评论必然没有其起到它原本应有的左右，诸如“很好”、“很喜欢”、“很棒”等。为此，我们需要对评论文本删除掉过短的文本信息，以去除那些对实验结果无用的信息。

在本数据集当中，文本字符长度低于等于 3 的有 65 条数据，将其剔除后还保留 44550 条评论数据。

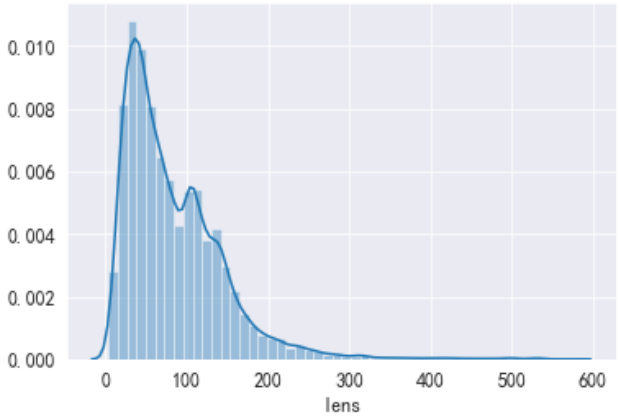


图 3.5 字符长度分布

对于评论文本的字符长度进行统计，字符长度均值为 83，字符最短为 4 个字符（删除 3 字符及以下），字符最长的评论为 576 个字符，可以看到大部分的文本字符长度都在 100 个字符以内。

```
count    44550.000000
mean        83.717598
std         61.664336
min          4.000000
25%         39.000000
50%         68.000000
75%        115.000000
max        576.000000
Name: lens, dtype: float64
```

图 3.6 字符长度描述统计

3.3 分词处理

正如前文所阐述的那样，对于数值型数据，每样本都具有很显著的数字特征，这些特征会包含各种类型的数值型数据，包括定距数据、定比数据等，对于这种类型的数据，能够很方便的进行数据插补、异常值检测等预处理，然后可以利用这些数据建立模型。但是本文的文本评论数据属于非结构化的数据，没有办法直接抽取其特征进行数据分析、建模，需要进一步的处理。

在中文文本当中，只有字、句、段具有很明显的分隔符进行切分，但是对于

一段文本当中的词和词组来说，它们之间的边界十分模糊，并没有一个非常明显的边界划分，所以在建立模型之前，需要对中文文本分词，文本分词的效果深刻地影响着后续建模的有效性，如果分词效果不尽人意，那么模型的结果也必定会产生很大的偏差。

在中文的自然语言处理领域，文本分词技术已经十分的成熟，可以很好的对文本信息进行切分，并且有着比较高的精确性。

本文采用 python 的中文分词工具‘jieba’对评论文本在去除停用词的基础上进行分词处理，部分样本的分词结果如图 3.7 所示。

外观 很惊艳 很难 相信 这是 几百元 低端 手机 屏幕 效果 心事 很 好 很 细腻 拍照 效果 不错 运行 速度 快 待机时间 长 尤其 5000 毫安 大 电池 续航 时间 长 不用
手机 外观 精致 大气 端庄 稳 稳妥 妥 蓝绸 摩擦 感 内有 5000mAh 电池 支持 18 瓦快 充 超 方便 Type C 接口 好 几天 才 充电 6.22 英寸 大屏 刚 贴膜 没 多久 起
帮人 一下 看 上 颜色 色彩鲜艳 十分 漂亮 MIUI 不是 盖 运行 流畅 5000 毫安 大 电池 配合 cpu 相当 省 电 保证 长时间 待机 操作 灵敏 不卡顿 不玩 大型 游戏 足
买 婆婆 高兴 坏 之前 他用 手机 内存 太小 卡 买个 新 内存 够用 款 不错 字体 很大 非常适合 老人家 很 方便 买 蓝色 耐看 好看 一段时间 很 适应 手机 价格 实惠
10 天 之内 买 2 台 几百元 来说 手机 只能 说千值 万值 屏幕 大 很 清晰 运行 速度 快 外形 漂亮 最 重要 5000 毫安 大 电池 确实 耐用 对比 很多 款 同价位 手机 -
宝贝 已经 收到 东西 非常 不错 性价比 很 高 起来 很 流畅 5 毫安 真的 很大 配合 处理器 达到 极限 省 电 超长 续航 后盖 手感 真的 很 好 好 舒服 屏幕 很 上 挡
外形 外观 外观 小巧 精致 塑料 质感 很 好 很 舒服 屏幕 音效 屏幕 价位 还 外放 很 大声 拍照 效果 扫个 码 水平 运行 速度 速度 很快 不卡 非常 满意 待机时间
屏幕 音效 屏幕显示 清晰 色彩 亮丽 音效 不错 拍照 效果 反应速度 很快 拍照 很 清晰 老人 买 看起来 很 不错 试一试 续航 时间

图 3.7 部分分词结果

从图 3.7 我们可以看到，jieba 分词的结果是十分精确的，在进行分词之后，文本中的一些没有实际意义的词汇都会被删除掉，比如“吗”、“啊”、“呀”等，这些被统称为“停用词”，停用词具有两个方面的特征：

（1）词汇极其普通，并且出现的频率非常高

（2）包含的信息量低，甚至没有任何信息，对文本的理解没有任何帮助，并且停用词的出现会对文本建模产生很大的影响。

文本所选用的部分停用词如表 3.5 所示。

表 3.7 部分停用词

部分停用词						
得	的	的话	等	等等	地	第
哈	对	对于	多	多少	而	而况
而且	而是	而外	而言	而已	尔后	反过来

本文经过 jieba 分词之后，结果示例如下所示：

手机 很 不错 使用 起来 很 流畅 是 一部 很 满意 的 手机

手机 不错 使用 起来 流畅 一部 满意 手机

从第一个示例可以看到，一般的文本评论当中包含着很多的介词或者副词作为语句的修饰词，这些词汇的存在与否并不会对文本的最终结果产生很大的影响，

反而这些没有任何意义的词汇的存在会造成一定的噪声，在词频列表当中占有相当大的比重，所以为了避免因为这些词汇的出现对分析结果产生不好的影响，所以需要加入停用词典，保留下对文本的理解有意义的词汇。

在去除停用词之后，由于中文语言博大精深，语义理解十分复杂，原生的 jieba 分词结果可能并不是特别理想，一些专有名词可能其无法理解，jieba 将其切分的粒度特别小，但是这些专有词汇对结果很重要，所以需要为其加入自定义词典，让其切词更加符合应用场景。

分词之后根据分词文本，绘制整体文本评论的词云图，如图 3.5 所示：

图 3.8　词云图

从词云图显示的结果我们可以看到，外观、屏幕、速度、拍照等词语出现的频率比较高，说明消费者对手机的这些特点关注度比较高。

经过文本分词之后，将初始的文本数据处理成为词汇的集合，方便进行进一步的向量转换。

3.4 文本向量化处理

在上一节，我们已经将评论变成有一个个词汇组成的评论语料文本，接下来就是要将这些词语转化成为向量来表示这些词汇。再将词汇转换成为词向量的过程当中，我们必须保证语义相近的词语，它们转换成为向量之后的距离也十分接近，本文主要采用分布式的词向量表示方法，更多的关于词向量表示的介绍和描

述请参照前文，本节将直接进行词向量的训练。

3.4.1word2vec 词向量化

本文采用分布式表示方法中 Word2vec 工具对词向量进行训练，Word2vec 由 skip-gram 和 cbow 这两个重要的模型组成，它们都包含了输入层、隐藏层和输出层，有所不同的是 skip-gram 是基于中心词来预测上下文，图 3.6 展现 Skip-gram 的基本思想。

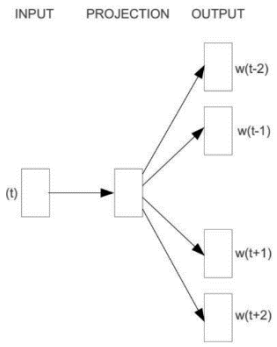


图 3.9 skip-gram 模型

Skip-gram 的目标函数与 CBOW 模型的目标函数基本类似，都是为了优化词向量矩阵 L 以最大化所有上下文词的对数似然：

$$L^* = \underset{L}{argmax} \sum_{w_i \in V} \sum_{w_c \in Wc} logP(w_c|w_i)$$

上式中，V 表示词表的大小，w_i表示中心词，w_c表示窗口为 c 的上下文背景词。

本文之主要采用 Skip-gram 模型对文本数据进行训练词向量，借助于 python 的 gensim 库开始进行词向量的训练，在训练词向量时，通过负采样的方式提高计算效率，一些有必要的参数设置如下所示：

```
In [90]: model = word2vec.Word2Vec(sentences, sg=1, size=100,  window=5,  min_count=3,  negative=3,
sample=0.001, hs=1, workers=4)
```

图 3.10 word2vec 参数设置

表 3.8 word2vec 参数设置表

参数	说明
Sentences	经过 jieba 分词之后所形成的文本数据
Size	希望经过训练之后得到的词向量的维度，本文所设置的 size 大小为 100，即每个一个词汇都由一个维度为 100 的向量表示
Sg	Sg 为模式参数，由 1 和 0 这两个选择，分别表示 skip-gram 和 CBOW 这两种模型，本文选择 1，也即 skip-gram 模型
Window	表示取样的窗口大小，本文设置为 3，表示每个样本都会对前 3 个词和后 3 个词进行取样
Min_count	对低频词过滤，本文设置为 5
Workers	并行化的数值，本文设置为 4

按照以上的参数进行训练词向量，并对词汇进行向量化的操作。如“续航”一词，将其向量化可以得到一个 100 维的向量，部分向量元素如下所示：

[1.82162262e-02, 7.42328390e-02, 1.71508536e-01, -2.99021542e-01, -1.70021772e-01, -5.56398481e-02, 2.48962551e-01, 4.06040102e-02, 4.43589509e-01, 2.58428901e-01, 1.97017685e-01, 5.41388169e-02,…,-2.14561909e-01, -3.72920066e-01, 3.47949229e-02, 4.51365232e-01, 3.91113368e-04, 1.18934959e-01, 8.67608339e-02, -8.65708068e-02]

将词语进行分布式表示，将其转换成为向量的目的就是考虑到文本之间语义的相似性，语义相近的词汇，希望其向量之间的距离也十分小。取出三个词语，分别为“续航”、“电池”、“声音”比较它们三者直接的距离。

表 3.9 相似度比较

词语	相似度
续航、电池	0.69
续航、音量	0.20
电池、音量	0.29

由表 3.9 可以看到，续航和电池的相似度很高，相似度约为 0.69，续航和音量之间的相似度为 0.20、电池和音量之间的相似度为 0.29。取出与续航相似

度最高的词语，如表 3.8 所示：

表 3.10 最高相似度

词语	相似度
蓄电	0.70
电池	0.69
大容量	0.63
充能	0.61
毫安	0.60
效能	0.60
待机	0.59
4500mA	0.59
待机时间	0.59

根据上述数据可以看到，经过训练，word2vec 很好的完成了我们的既定目标，考虑到了文本语义之间的相似性，语义越相近的词语，其向量之间的距离越是接近。在将词语转化成向量之后，还仍需要将句子也转换成为固定维度的向量表达。

3.4.2Tf-IDF 向量化

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。将文本 TF-IDF 词向量化，如下例所示：

表 3.11 if-idf 向量化

手机	很不错	外观	很	好看	运行	速度
0.056	0.079	0.23	0.007	0.45	0.51	0.68
特别	快	听歌	时候	音质	很	悦耳
0.06	0.001	0.14	0.069	0.368	0.057	0.685

将其作为文本特征输入到模型之中，进行模型的训练。

3.5 句向量的构造

上一节主要讲述了如何训练词向量，将词汇转换成为 100 维的数值向量，接下来需要进一步进行句子向量的构建。本文采用将每个样本之中，相似的语义词

向量取平均，再将其与其他向量相加，得到一个长度为100维向量，以这个向量来表示这个句子样本。

首先，计算每条样本中的词向量相似度，也即词向量之间的距离，然后设置阈值，如果向量之间的相似度高于这个阈值，则说明这两个词向量之间具有相似性，用它们的平均值来表示这几个词汇。

假设样本中有 N 个词语的相似度高于阈值，则它们的均值向量v_s如下所示：

$$v_s = \frac{\sum_{i=1}^{N} v_i}{N}, i = 1,2,\ldots,N$$

其中，v_i为通过 Word2vec 训练出来的第 i 个词向量，v_s即为这 N 个相似词向量的均值词向量。

假设某个样本共有 M 个词汇，共有 N 个相似词向量集合，v_{si}表示第 i 个相似词向量集合的均值，集合中有N_i个相似的词语，则该样本的句子向量如下所示：

$$S_j = \sum_{i=1}^{N} v_{si} \times N_i + \sum_{j=1}^{M-N} v_j$$

其中，v_j表示普通的词向量（没有与之语义相近的词）。

本章主要介绍文本的分词以及词向量、句向量的构建。首先，对文本数据进行清洗，在清洗过后对文本分词，将句子切分成若干个词汇，再对文本进行词向量的训练。在训练出来的词向量的基础上，对文本进行构造句向量，以这个句子向量来表示整个句子。

3.6 数据标注

3.6.1 相似评论的标注

相似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阈值(0.9)，则将该句子标记为重复评论，并将其归入标记数据集的候选评论。

计算评论相似度的过程如下：

(1)数据清洗；

(2)考虑到上下文语义关系，构建评论的语料库，利用 word2vec 将每条评论转换为句向量；

(3)计算文本与文本之间相似度（余弦相似度）

根据评论间的相似度，取出相似度大于阈值(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。

3.6.2 评论时间异常的标注

3.6.2 数据标注

在经过前文的处理之后，本共 44550 条文本评论数据，接下来需要准备训练数据集，并对训练数据集进行标注。

数据标注需要耗费大量的人力，从成本的角度出发，本次研究从每家店铺随机抽取 1666 条文本评论，组成本次研究的训练集，共计 10000 条文本需要标注。本次研究组织了五名电子商务专业的研究生，在熟悉标注规则的条件下，随文本评论进行标注，在标注的过程当中，会综合考虑前文提及的一些因素（各家店铺评论数量异常日期、重复评论等），对并从虚假评论高重复、时间点异常、是否包含其他平台名称、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，这些完成标注的数据将作为后续建模过程的训练数据。

如下展示两条专家标记为虚假评论的样例：

（1） 手机很不错，性价比很高，漂亮大气，很是推荐大家购买，买不了吃亏，买不了上当，物流很快。

（2） 和在专卖店上面买的一样，比在**上面买的好多了，信赖**，客服很有耐心，还会再来的。

4. 虚假评论识别

4.1 基分类器的构建

4.1.1 模型的评价方法

本文主要使用有监督的方法对京东商城的评论数据进行训练并、识别虚假评论,在本质上,它属于有监督的二分类任务,故本文主要使用精确度(precision)、召回率（Recall）和 F1-score 对模型的结果进行评价。

首先得出模型结果的混淆矩阵，如表 4.1 所示：

表 4.1 混淆矩阵

		预测结果	
		正（P）	负（N）
真实结果	正（P）	TP	FN
	负（N）	FP	TN

其中，TP（真正）表示模型预测结果为正，样本的真实结果也为正；FP（假正）表示模型预测结果为正，真实结果为负；TN（真负）表示模型预测结果为负，真实结果也为负，FN 指的是模型预测结果为负，真实结果为正。

TP + FN + FP + TN = 样本总数

$$精确率 = \frac{TP}{TP + FP}$$

$$召回率 = \frac{TP}{TP + FN}$$

一般而言，为了提高模型的性能分类器需要提高精确性，那么它为了模型的可靠性，会放弃一些没有把握的样本。这个时候就需要另外的指标来做一些平衡，综合考虑精确率和召回率，F1-score 就是这样一个指标，它是精确率和召回率的调和平均值：

$$\frac{2}{F_1} = \frac{1}{P} + \frac{1}{R}$$

4.1.2 数据集的划分

在对数据进行标注之后，共得到 7836 条真实评论，2164 条虚假评论。可以看到数据存在这严重的不平衡，为了减轻数据不平衡对后续建模产生的影响，本文采用 bootstrap 抽样的方法从 0 类数据从抽取 4000 条将其放入到训练数据当中，最终可以得到真实评论 7836，虚假评论 6164 条，共计 14000 条评论数据。另外，将标注好的数据集按照 8：1：1 的方式划分为训练集、验证集、测试集，来进行模型的训练和测试。

为了充分比较各模型组合之间的差异，本文分别采用 TF-IDF、Word2vec 两种词向量化的方法进行模型的训练。

4.1.3Logistic Regression 模型

Logistic Regression 能够很好地针对二分类问题对问题进行建模分析，他被应用在很多领域当中，例如广告的点击率、垃圾邮件的分类、病人患病情况诊断、金融诈骗、虚假账号检测等领域。Logistic Regression 采用对数似然作为其损失函数，使用梯度下降法进行模型的求解，与此同时它具有多个局部最小值，尽管它没有全局最优值，但是局部最优值的效果也很不错。

首先利用 logistic Regression 对训练数据进行训练，并且为模型增加正则项，防止过拟合的现象出现。

提取训练数据的句子特征并进行向量操作，将利用 word2vec 训练好的词向量输入到模型当中，来进行模型训练。随后根据结果相关的评价指标选择合适的参数，完成上述步骤之后得到最优模型并保存。

模型训练过程如图所示，可以看到随着正则参数 c 的变化，模型的效果也在发生相应的变化。

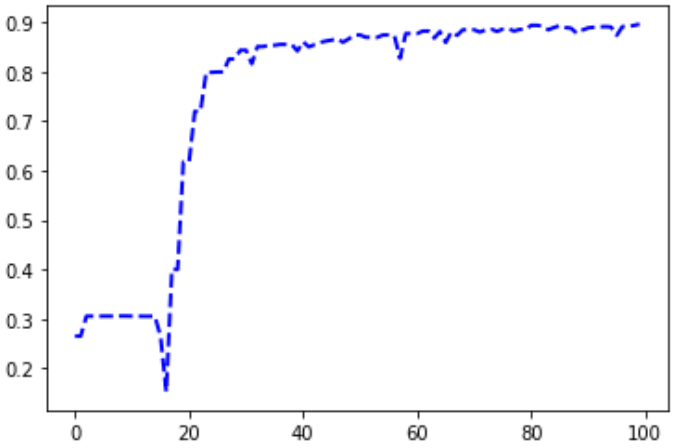


图 4.1 训练过程

根据精确率、召回率、F1 得分综合考虑，选择最优参数并保存最优的模型，对测试数据进行测试，得到混淆矩阵如表 4.2 所示：

表 4.2 混淆矩阵

Log+word2vec		预测结果	
		正（P）	负（N）
真实结果	正（P）	619	165
	负（N）	99	517

Word2vec+logistic Regression 方法在测试集上的分类精确率为 0.861，召回率为 0.794，F1-score 为 0.826。

使用 if-idf 方法对文本进行词向量的转化，从 python 的 sklearn 库中导入相关的词向量化的包，from sklearn.feature_extraction.text import CountVectorizer, TfidfTransformer，将分词后的评论文本数据作为参数传入，再利用训练数据训练模型，随后根据结果相关的评价指标选择合适的参数，完成上述步骤之后得到最优模型并保存。

表 4.3 混淆矩阵

log+if-idf		预测结果	
		正（P）	负（N）
真实结果	正（P）	576	208
	负（N）	178	438

tf-idf+logistic Regression 方法在测试集上的分类精确率为 0.763，召回率为 0.734，F1-score 为 0.748。

4.1.4 支持向量机分类器

支持向量机作为监督学习领域特别经典的分类器被人们所熟知，因为它将问题转化成为一个凸二次规划的问题，所以它与 logistic Regression 不同，它具有全局最优解。

提取训练数据的句子特征并进行向量操作，输入到模型当中，来进行模型训练。随后根据结果相关的评价指标选择合适的参数，完成上述步骤之后，得到训练好的最优模型。

本文使用径向基核函数（RBF）进行构造 SVM 模型，提取训练数据的句子特征并进行向量操作，输入到模型当中，来进行模型训练。随后根据结果相关的评价指标选择合适的参数，得到训练好的最优模型。

根据精确率、召回率、F1 得分综合考虑，选择最优参数并保存最优的模型，并使用测试数据进行预测并记录其结果，得到混淆矩阵如表 4.4 所示：

表 4.4 混淆矩阵

svm+word2vec		预测结果	
		正（P）	负（N）
真实结果	正（P）	576	208
	负（N）	249	367

在测试集上的分类精确率为 0.698，召回率为 0.735，F1-score 为 0.716。

使用 if-idf 方法对文本进行词向量的转化，将分词后的评论文本数据作为参数传入，再利用训练数据训练模型，随后根据结果相关的评价指标选择合适的参数，完成上述步骤之后得到最优模型并保存。

模型训练过程如图所示，可以看到随着参数 gamma 的变化，模型的效果也在发生相应的变化。

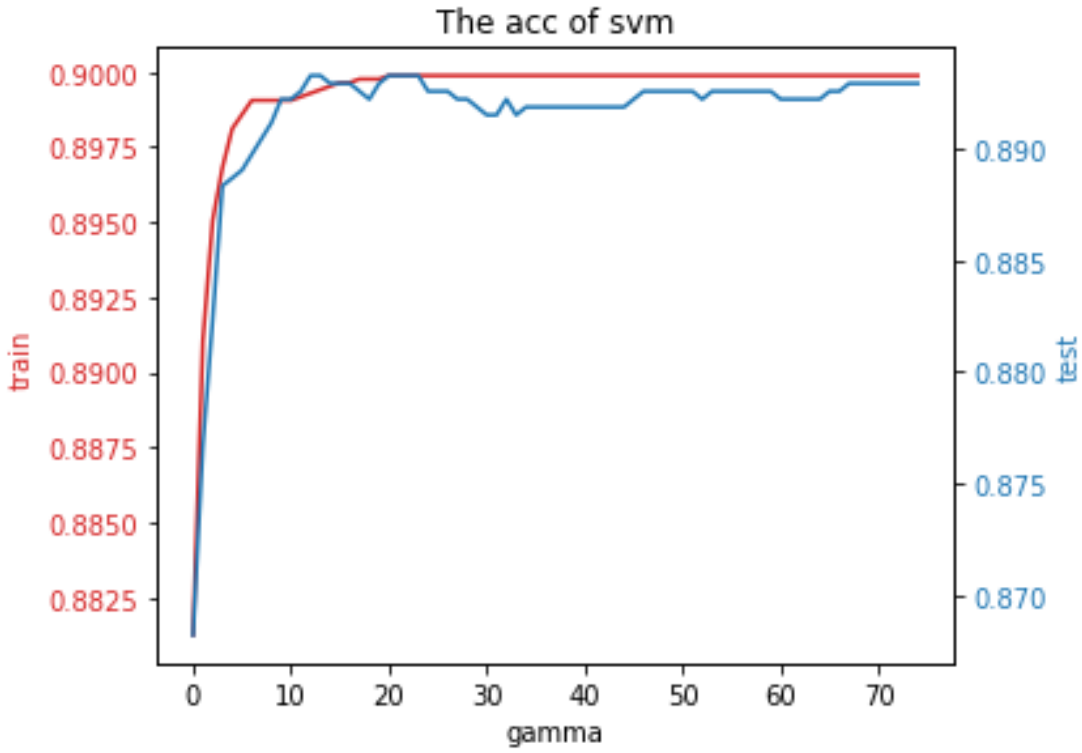


图 4.2 训练过程

选择最优参数并保存最优的模型，并使用测试数据进行预测并记录其结果，得到混淆矩阵如表 4.5 所示：

表 4.5 混淆矩阵

Svm+tf-idf		预测结果	
		正（P）	负（N）
真实结果	正（P）	638	146
	负（N）	72	544

在测试集上的分类精确率为 0.898，召回率为 0.813，F1-score 为 0.854。

4.1.5 朴素贝叶斯分类器

朴素贝叶斯常用于文本分析领域，最擅长的领域就是文本分析，包括文本分类、情感分析、垃圾邮件处理等。

本文采用多项朴素贝叶斯分类器，针对评论文本进行分类。首先，导入必要的库和包，from sklearn.naive_bayes import MultinomialNB，提取训练数据的句子特征并进行向量操作，将文本评论数据向量输入到模型当中，来进行模型

训练。随后根据结果相关的评价指标选择合适的参数，完成上述步骤之后得到最优模型并保存。

选择最优参数并保存最优的模型，并使用测试数据进行预测并记录其结果，得到混淆矩阵如表 4.6 所示：

表 4.6 混淆矩阵

bayes+word2vec		预测结果	
		正（P）	负（N）
真实结果	正（P）	503	281
	负（N）	348	268

在测试集上的分类精确率为 0.591，召回率为 0.641，F1-score 为 0.615。

使用 if-idf 方法对文本进行词向量的转化，将分词后的评论文本数据作为参数传入，再利用训练数据训练模型，随后根据结果相关的评价指标选择合适的参数，完成上述步骤之后得到最优模型并保存。

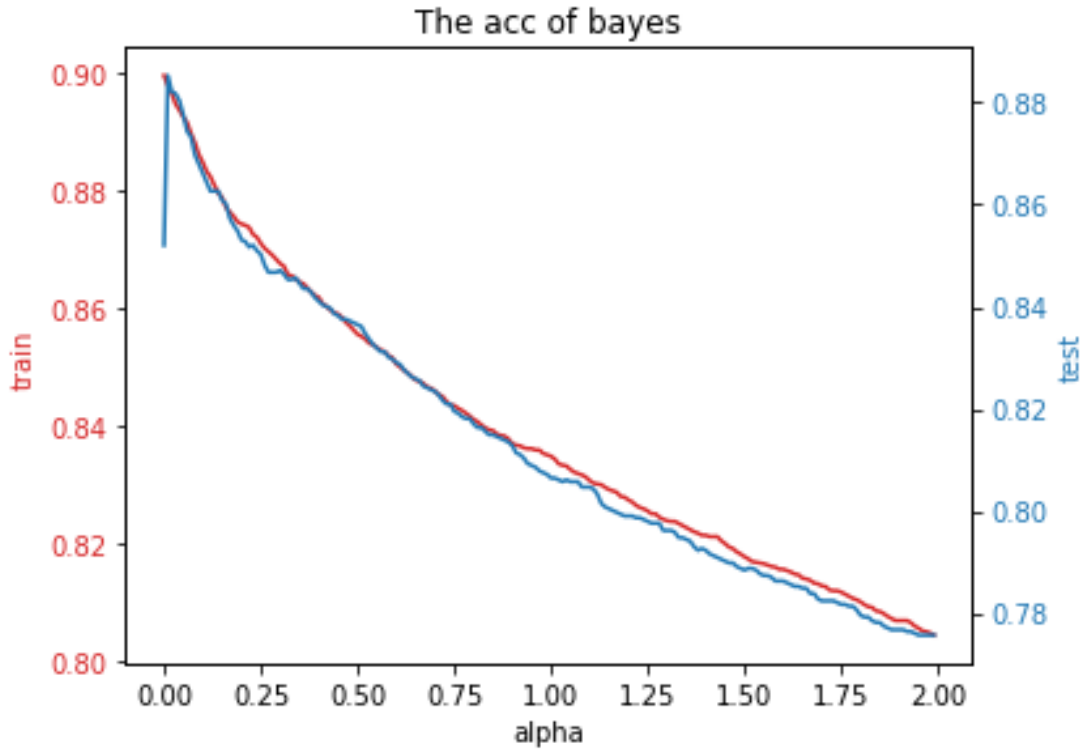


图 4.3 训练过程

选择最优参数并保存最优的模型，并使用测试数据进行预测并记录其结果，得到混淆矩阵如表 4.7 所示：

表 4.7 混淆矩阵

bayes+tf-idf		预测结果	
		正（P）	负（N）
真实结果	正（P）	633	151
	负（N）	82	534

在测试集上的分类精确率为 0.885，召回率为 0.807，F1-score 为 0.845。

前面，本文通过两种文本特征向量化的方式进行建模、预测、评估，可以看到在本文当中，每个分类器的效果都有着显著的不同，分类的具体效果如表 4.8 所示：

表 4.8 模型对比

分类器	精确率	召回率	F1-score
$Logstic + tf\_idf$	0.763	0.734	0.748
$Logstic + word2vec$	0.861	0.794	0.826
$SVM + tf\_idf$	0.898	0.813	0.853
$SVM + word2vec$	0.698	0.735	0.716
$Bayes + tf\_idf$	0.885	0.801	0.845
$Bayes + word2vec$	0.591	0.641	0.615

由表 4.9 可以清晰的看到，各个特征化方法、各个分类器在本文数据集当中的表现。在 logistic regression 方法当中应用 tf-idf 的特征化方法，其精确率为 0.763，召回率为 0.734，F1-score 为 0.748，而应用 word2vec 特征化方法其精确率为 0.861，召回率为 0.794，F-score 为 0.826，可以看到在本文场景当中，对于 logistic regression 而言选择 word2vec 的方法更加适合；对于支持向量机(SVM)分类器而言，使用 tf-idf 方法的精确率为 0.898,召回率为 0.813，F1-score 为 0.853,而应用 word2vec 的方法,其精确率为 0.698,召回率为 0.735，F1-score 为 0.716，应用 tf-idf 的特征化方法更加适合；对于朴素贝叶斯分类器而言，应用 tf-idf 特征化方法，其精确率为 0.885，召回率为 0.801，F1-score 为 0.845，而应用 word2vec 特征化方法，其精确率为 0.591，召回率为 0.641，F1-score 为 0.615，其性能均要低于传统的 tf-idf 方法。

4.1.6 基于 Text-CNN 模型的分类器

Text-CNN 模型的输入需要固定输入序列的长度，比这个长度短的文本序列

需要进行填充（padding）操作，比这个固定长度长的文本序列就需要进行截断操作，以期达到每个样本序列等长的目标，根据前文所分析的那样，语料的平均长度为 100 个字符，本文设置定长为 100，最终输入层的输入为文本序列当中各个词汇对应的 word2vec 词向量。

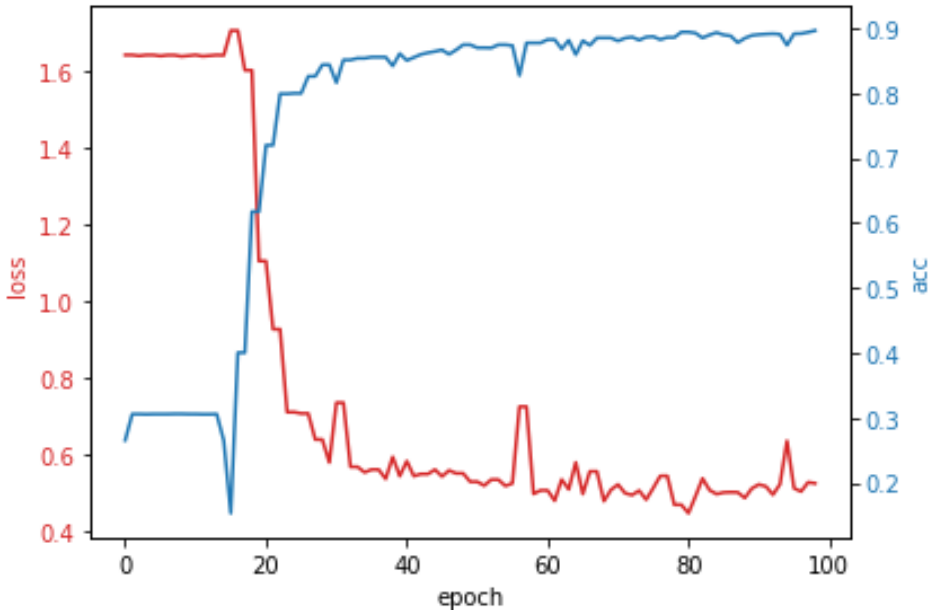
在 Text-CNN 当中，卷积核的高度可以理解成为 N 元语言模型当中的 N，一般都会使用不同大小的卷积核，卷积核的高度也被称之为窗口值，一般取 2-8 之间的数值。

本文采用的优化算法为 adam 算法，相关的参数设置如表 4.9 所示：

表 4.9 参数设置

参数	取值
Batch_size	256
N_epoch	50
Learning_rate	0.001
L2_reg_lambda	1.00E-08
Embedding_size	100
Num_class	2
Sequence_length	100

其中，batch_size 表示批次大小；n_epoch 表示迭代的次数为 20；learning_rate 表示学习率；L2_reg_lambda 表示正则化处理时的正则化参数；embedding_size 表示词向量的维度；num_class 表示输出层的维度，设置为 2。



对训练数据进行训练，每个迭代周期进行一次验证集上的验证，训练过程由图 4.4 所示。图 4.4 为模型在验证集上的损失表现情况，横轴为迭代次数，可以看到模型的损失处于波动下降的状态，在近 60 次迭代后不再有明显的下降情况。模型在近 60 次迭代后，模型的准确率趋于稳定，准确率稳定在 85%和 90%之间，模型收敛，根据精确率和召回率的情况保留最优的模型参数。

选择最优参数并保存最优的模型，并使用测试数据进行预测并记录其结果，得到混淆矩阵如表 4.10 所示：

表 4.10 混淆矩阵

Text-CNN		预测结果	
		正（P）	负（N）
真实结果	正（P）	645	139
	负（N）	87	529

如表 4.10 所示，使用 text-cnn 保存的最优模型对测试集进行测试，其精确率为 0.881，召回率为 0.822，F1-score 为 0.853。

4.1.7 基于长短期记忆网络(LSTM)的分类器

本文采用双向长短期记忆网络，采取 drop_out 的方法减缓过拟合的问题。Drop_out 就是按照一定的概率取去除一些神经网络的神经元，在前向传播的过程当中，只是激活所有神经元当中的部分神经元，每次训练都会随机激活某些神经元，这样的方法在一定程度上缓解了 LSTM 当中的过拟合问题。

表 4.11 参数设置

参数	取值
Epoch	160
Embedding_size	100
class	2
Seq_len	100
Drop_out	0.4
Lr	0.001
Batch_size	128
Num_layer	2
Grad_clip	20

对 LSTM 进行训练，参数主要有输入向量的维度，隐藏层的数量，学习率，词典大小等，相关的主要参数设置如表 4.11 所示。

表 4.11 中 epoch 表示模型的迭代次数，本文设定为 100 次；embedding_size 代表词向量的维度；class 指的是模型的输出层的维度；seq_len 指的是本条样本的长度；drop_out 为缓解过拟合的措施，本文设定为 0.4；lr 为学习率；batch_size 为每个训练批次的样本量大小；num_layer 为神经网络的层数本文为双向双层 LSTM，故设定为 2；grad_clip 为梯度裁剪的阈值，因为循环神经网络会对历史信息进行记忆，容易产生梯度爆炸，所以需要进行梯度裁剪，本文设置为 20。

根据以上的参数对训练数据进行训练，训练过程中，每个迭代周期都要对验证数据进行验证并记录下来，如图 4.6 所示：

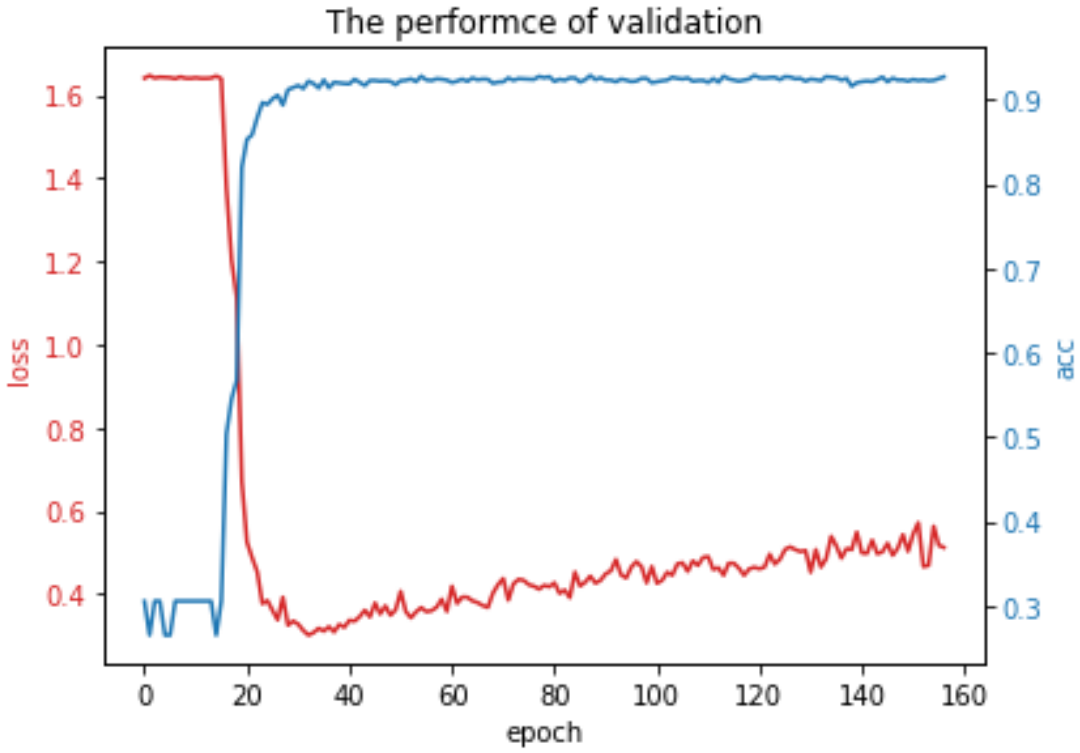


图 4.6 训练过程

图 4.6 为 BiLSTM 模型的训练过程，入图所示，随着迭代次数的增加，验证集损失呈现出降低的趋势，于此同时，验证集的准确率也在不断上升，模型在迭代次数带到 40 次，准确率达到 90%左右收敛，并保留最优的模型参数。

选择最优参数并保存最优的模型，并使用测试数据进行预测并记录其结果，

得到混淆矩阵如表 4.12 所示：

表 4.12 混淆矩阵

lstm		预测结果	
		正（P）	负（N）
真实结果	正（P）	667	117
	负（N）	59	557

如表 4.12 混淆矩阵所示，用保留的最优模型对测试数据进行预测并验证，其精确率为 0.919，召回率为 0.850，F1-score 为 0.883。

4.2 模型融合

为了进一步提升模型的效果和提升模型的稳定性，本文借鉴集成学习的方法，对上诉模型进行融合。一般的模型集成采用类似于 adaboost 的方法，它们会根据每次迭代的结果来更新样本的权值，利用更新的权值来进行下一次迭代，但是对于这种方法，它们的基分类器或者处理特征的方法一般都是相同的，本文经过前文的实验比较，在本数据集中，由于处理特征的方法不是完全相同，所以不能按照以往的集成方法来进行模型融合。本文主要采取加权投票法来进行模型的融合，让性能更好的模型能够掌握更多的投票权，这样能够最大程度上地纠正单个分类器的初始分类错误。

按照以上的思想，构建加法模型，得到 5 个基本类器的误分率和权重系数如表 4.13 所示：

表 4.13 正例组合模型

模型	准确率	权重系数
$logistic\ regression + word2vec --G_1$	0.86106	0.193648987
$SVM + tf\_idf --G_2$	0.89781	0.201913889
Naive Bayesia + tf_idf $--G_3$	0.886239	0.199311463
Text CNN + $word2vec --G_4$	0.88225	0.19841436
LSTM + $word2vec --G_5$	0.919142	0.206711301

所以，关于正例组合模型为：

$$G_{+}(x) = 0.193G_1(x) + 0.202G_2(x) + 0.199G_3(x) + 0.198G_4(x) + 0.206G_5(x)$$

表 4.14 负例组合模型

模型	准确率	权重系数
$logistic\ regression + word2vec --G_1$	0.839975	0.192974
$SVM + tf\_idf --G_2$	0.883371	0.202943
Naive Bayesia + tf_idf $--G_3$	0.866497	0.199067
Text CNN + $word2vec --G_4$	0.858733	0.197283
LSTM + $word2vec --G_5$	0.904220	0.207733

所以，关于负例的组合模型为：

$$G_{-}(x) = 0.193G_1(x) + 0.203G_2(x) + 0.199G_3(x) + 0.197G_4(x) + 0.208G_5(x)$$

利用以上的组合模型在测试集上进行预测，并得到关于测试数据的混淆矩阵如表 4.15 所示：

表 4.15 混淆矩阵

融合模型		预测结果	
		正（P）	负（N）
真实结果	正（P）	675	109
	负（N）	47	569

其精确率为 0.934，召回率为 0.861，F1-score 为 0.896。

表 4.16 模型比较

分类器	精确率	召回率	F1-score
$logistic\ regression + word2vec$	0.861	0.794	0.826
$SVM + tf\_idf$	0.898	0.813	0.853
Naive Bayesia + tf_idf	0.885	0.801	0.845
Text CNN + $word2vec$	0.882	0.822	0.851
LSTM + $word2vec$	0.919	0.850	0.883
融合模型	0.934	0.861	0.896

如表 4.16 所示，可以看到经过模型融合之后得到的模型精确率、召回率和 F1-score 得分都有一定程度的提升，其结果都要优于任何一个单一的基分类器，

模型的结果更加稳定。

利用得到的组合模型对剩下的 34550 条未标注数据进行虚假评论的预测，得出所有的样本标签，以便进行下一步的数据分析，模型预测结果如表 4.17 所示：

表 4.17 模型预测

标签	数量
虚假评论	9167
真实评论	25383

可以看到，通过组合模型对文本评论数据进行预测，有 9167 条评论数据为虚假评，占总体的 26.53%，有 25383 条评论为真实评论，占总体的 73.47%。

5. 虚假评论模式分析

5.1 评论特征提取

前文已经根据训练所得出的比较稳定的组合模型对京东商城的评论文本进行虚假评论的识别，而且得出了所有的样本的标签值，接下来需要对虚假评论的特征词进行提取，本文主要通过 tf_idf 对两类文本进行语言特征的提取，查看 tf_idf 高的值，在这两类样本当中有什么显著的区别，比较二者之间的差异性。

前文已经对 tf_idf 进行简单的介绍，TF 指的是词频，就是某个词语出现的次数，某些常见的词出现的概率会比较高，比如“的”、“很”等，这些词对分析的结果并不能产生正面的影响，反而会增加一些噪声。所以，为了避免这些高频词对结果产生不好的影响，有学者提出了 idf，也叫逆文档频率，对 TF 进行修正。

对于所有的虚假评论进行特征提取，共抽取了 788 个特征词语，由于维度较高，所以 tf_idf 矩阵较为稀疏，如下所示：

```
[(0, 5.892391025913404),
 (1, 4.913707257150496),
 (2, 1.7191569670278068),
 (3, 7.068098379949722),
 (4, 10.443137811296648),
 (5, 5.582050905301253),
 (6, 7.554394562398388),
 (7, 9.398743691938193),
 (8, 7.560494761934806),
 (9, 9.19521029785306),
 (10, 8.033746875158945),
 (11, 0.8530186371894716),
 (12, 12.121209716409286),
 (13, 2.7158554820844585),
 (14, 1.1806024707703644),
 (15, 6.021073045123835),
 (16, 1.0693890175938596),
 (17, 11.742698093155555),
 (18, 1.4577405603742206),
```

图 5.1 词索引及 tf-idf 值

本文以每个特征词 if-idf 的均值表示这个词语在文章中的重要性，得到每个词语及其 if-idf 值，虚假评论的主要特征词如下所示。

表 5.1 虚假评论主要 tf-idf 值

虚假评论主要特征词及其 tf_idf 值						
不错	手机	满意	客服	物流	质量	值得
0.091	0.087	0.083	0.081	0.074	0.071	0.064
购买	服务态度	喜欢	很快	贴心	京东	正品
0.051	0.048	0.045	0.041	0.041	0.039	0.035

对预测得到的虚假评论语料进行分析，得到 if_idf 最高的词语如表 5.2 所示：

表 5.2 真实评论主要 tf-idf 值

真实评论主要特征词及其 tf_idf 值						
手机	不错	流畅	屏幕	喜欢	满意	运行
0.088	0.088	0.087	0.078	0.077	0.071	0.069
续航	质量	喜欢	性价比	像素	收到	电量
0.064	0.053	0.051	0.049	0.034	0.028	0.02

从表 5.2 的两个表比较可以看到，虚假评论和真实评论想不，在某些特征词上比较相似，比如“手机”、“不错”、“满意”、“喜欢”、“质量”等。但是在虚假评论的权重比较高的词汇当中，包含“客服”、“物流”、“值得”、“购买”、“服务态度”、“正品”等，这些词汇多集中于物流、服务等方面，对商品本身的描述词与比较少，而对于真实评论来说特征词主要集中在“屏幕”、“流畅”、“运行”、“流畅”,这些词主要都集中在对商品的实物描述上,两类评论文本展现出很明显的不同,由于后续的特征词的 tf_idf 值比较小，所以没有继续进行展示，但是通过粗略的观察可以看到，在真实评论当中，“颜色”、“外观”、“速度”、“拍照”等特征词的重要程度较高，然而这些词语在虚假评论当中的比重比较低。可以看到真实评论由于是消费者真实使用体验的产品,所以消费者所发布的评价信息大都聚焦于商品本身的细节，会给出一些自己真实的感受和具有针对性的评价。但是虚假评论的模式主要是一些对于物流、客服等的评价，对商品本身的细节、使用感受描述比重比较低。

5.2 语言模型差异

根据前文的描述，不难知道 tf-idf 属于词袋模型，他有着自己的缺点，那就是词袋模型不会考虑到上下文语义化之间的关联性，但是评论语料之中肯定包含了语序信息，所以，本文在考虑 tf-idf 进行特征词汇抽取之外，还需要考虑文本的语序信息，构建文本评论的语法模型，以此来挖掘文本的前后文关系和语序信息。

语言模型，就是判定一句话是否出现的概率。假设一句话由 n 个词语组成，那么语言模型就可以以这样的形式来表示一句话出现的概率：

$$P(S) = P(W_1, W_2, \dots, W_n) = P(W_1)P(W_2|W_1) \dots P(W_n|W_1, W_2, \dots, W_{n-1})$$

其中，W_i 表示句子中出现的词语。

但是这种表达方法存在着两个缺陷。首先，参数空间太大，条件概率的可能性太多，无法进行估算；其次，对于非常多词对的组合，如果它在语料库中没有出现，那么依照极大似然估计的原理，这个句子出现的概率为 0，这显然是不合理的。

为了解决上述问题，学者们引入了马尔可夫假设，随意一个词语出现的概率只与它前面出现的有限个或者几个词语有关。

如果一个词的出现只依赖于它前面的两个词语，就称之为三元语言模型。

$$P(S) = P(W_1, W_2, \dots, W_n) = P(W_1)P(W_2|W_1)P(W_3|W_1, W_2) \dots P(W_n|W_{n-1}, W_{n-2})$$

本文主要采用三元语言模型来比较虚假评论和真实评论之间的差异性。对于虚假评论，三元语言模型得到的几条词汇为：“手机 物流 很快”、“高端 大气 上档次”、“客服 服务 态度”、“京东 物流 正品”，而对于真实评论，得到的词语组合为：“手机 运行 流畅”、“屏幕 鲜艳 拍照”、“拍照 像素 喜欢”、“续航 电量 毫安”。

可以很明显的观察到两类评论之间的语言差异性，真实评论对商品的描述和特点更加的细致，囊括了产品的各个方面的使用感受；而虚假评论的描述比较空泛，比较突出体现在物流、客服等，对产品自身的描述比较少。

5.3 主题差异

本文通过使用 LDA 主题建模方法分别对两类文本进行 lda 主题分析，展现出两类评论在 5 个主题上的分布，分析两类评论在各自主题上的差异。

表 5.3 虚假评论 LDA 主题分布

主题	主题词
topic1	物流；很快； 速度；手机； 行；快递；发货；满意；服务
topic2	家人； 软件； 声音； 字体； 本来； 太小；运行； 赠品； 信号
topic3	客服； 京东； 申请； 性价比； 预装； 垃圾；质量；太卡；不好
topic4	品牌； 喜欢； 价格； 推荐； 父母； 帮；控制；使用；实用
topic5	物有所值； 手机； 喜欢； 买来； 膜； 合适；功能； 信赖； 实惠

表 5.4 非虚假评论 LDA 主题分布

主题	主题词
topic1	外观；颜色； 满意； 质量； 推荐； 性价比； 京东；大小；快递；好
topic2	流畅； 运行； 内存； 容量； 很好； 发热； 游戏；喜欢；卡
topic3	屏幕；质量； 物流； 指纹； 很快； 包装；字体； 清晰；京东
topic4	声音； 像素； 快； 拍照；品牌； 舒服； 性能； 特别
topic5	电量；微信；游戏；电话；很棒；时间；续航；服务；耗电

通过以上两张表的比较可以看到，两类评论在主题分布上存在着一些差别，如虚假评论当中存在着一些比较宽泛的词语，没有聚焦于商品具体的某个细节的评论比较少，大都是对物流、价格、客服、喜好、品牌等进行评价，反观非虚假评论，可以看到主题一的词汇有外观、颜色、大小等，不难发现是对手机的外观大小进行评价；主题二的主题词有流畅、运行、内存、容量、卡等词汇，这部分语料是对手机的性能进行评价的；主题三的主题词有屏幕、质量、字体等，主题四的主题词有声音、像素、拍照等，主题五有电量、时间、续航等，可以看到非虚假评论大都聚焦于商品各个方面的细节评价。

5.4 词性分析

词性标注是属于自然语言处理当中的基础步骤，是在给定句子中判定每个词最合适的词性标记，某个词属于名词、动词、形容词的过程。本文为了发掘文本

语料的有效信息，对两类已经标注好标签的文本评论进行词性标注。词性标注一般包含一级标注或者二级标注，一级标注一般只是显示单一的编码，用“n”表示名词，用“v”表示动词，二级标注更加具体，显示复合编码，比如副形词用“Ad”来表示，人名用“Nr”来表示，常用的词性表示方法如表 5.5 所示。

表 5.5 词性表示方法

词性编码	词性名称	词性编码	词性名称
n	名词	Ng	名语素
v	动词	Nr•	人名
a	形容词	Ns	地名
d	副词	Nt	机构团体
f	方位词	Vd	副动词
g	语素	Vg	动语素
i	成语	Ad	副形词
m	数词	An	名形词
t	时间词	Tg	时语素

本文借助 jieba 分词工具提供的词性标注的功能来对两类文本语料进行词性标注，于此同时对两类文本中的各词性出现的频次进行统计，以此来比较各词性在两类文本中的出现的差异，各词性在两类文本中出现的情况如表 5.6 所示。

表 5.6 词性比例

		虚假评论	非虚假评论	虚假比例	非虚假比例	差值
名词	n	2,331	4,572	0.15	0.18	0.04
动词	v	1,763	2,010	0.11	0.08	-0.03
形容词	a	979	3,356	0.06	0.14	0.07
副词	d	527	738	0.15	0.05	-0.10
方位词	f	337	780	0.02	0.03	0.01
数词	m	561	1,262	0.04	0.05	0.02
时间词	t	248	666	0.02	0.03	0.01

词性标注后的非虚假评论和虚假评论数据集的总体词性数量分别为 24751、15684，通过计算两个数据集当中不同的词性占据总体词性的比例，再比较二者大小，可以得到不同词性在不同的数据集中的占比情况。从表 5.6 可以看到，不论是虚假评论或者是非虚假评论，名词的数量都是最多的，虚假评论数据集当中，

名词的数量为3114个，将近占据总体的20%，而非虚假评论当中，名词的数量为6056个，占据总体的24%，占比高于虚假评论。其次是动词、形容词和副词等。另外，对不同的数据集各个词性占比进行差值计算可以看到，非虚假评论中名词的比例要高于虚假评论的比例，非虚假评论当中要有更多的对手机各个方面性能的介绍；非虚假评论中形容词的比重也要高于虚假评论，因为购买过商品的消费者会有更多的对商品具象的描述，运用更多的感觉上的描述，因此会更高频率的使用形容词。非虚假评论当中，数词和时间词比重也高于虚假评论，说明在虚假评论当中，对时间、数量大小等该类的描述语句比较少。

5.5 情感差异

伴随着数据挖掘和自然语言处理技术的发展，对于用户情感的分析也越来越凸显出它的价值，企业或者可以根据情感分析关注到消费者对于某个商品或者事件情感动向，以此来调整、制定相关的发展战略。对于本文来说，可以通过对于上述两类评论进行情感分析，以此来探究两类评论当中，正负情感的比重，为商家或者平台提供某种程度的参考。

本文的研究对象是京东平台的某手机品牌的在线评论，本文主要采用深度学习当中的LSTM对两类评论进行情感分析。情感分析的主题以及核心就在于“情感”，也即每条评论所代表的主观态度是怎么样的，评论者表达的是积极的情绪，还是消极的情绪。

首先，需要准备训练数据，需要大量的数据对模型进行训练。因为本文主要是电商平台的在线评论数据，为了保持与研究主体的相关性，本文收集了一万五千条有关电商领域的语料数据，涉及到家用电器、手机数码等领域的已标注评论数据。其中，正面情绪数据8600条，将其标记为“1”，负面情绪数据6400条，将其标记为“0”。并将这些数据按照7：2：1的比例划分为训练集、测试集、验证集。训练数据10500条，测试数据3000条，验证数据1500条。

与前文的数据预处理步骤相同，首先需要对文本数据进行预处理，包括数据清洗、分词等操作，随后利用word2vec工具对文本进行词向量化操作，并准备进行模型的训练。

模型的相关参数设置如下：

表 5.7 参数设置

参数	取值
Epoch	100
Embedding_size	100
class	2
Seq_len	100
Drop_out	0.5
Lr	0.01
Batch_size	64
Num_layer	1
Grad_clip	20

根据上述参数设置进行模型训练，图 5.1 为模型的训练过程，随着 loss 的下降，精确率也逐步提高到一个稳定的水平。并选取最优模型。

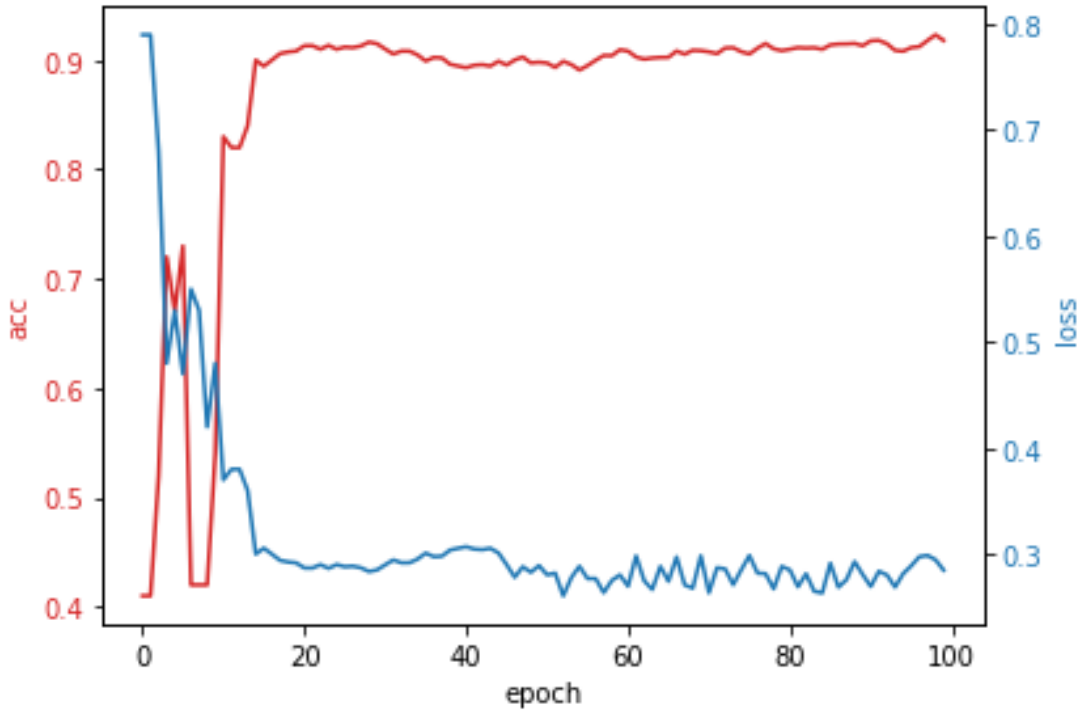


图 5.1 训练过程

利用测试集数据进行预测，得到混淆矩阵如下：

表 5.8 混淆矩阵

融合模型	预测结果	
	正（P）	负（N）

真实结果	正（P）	1528	192
	负（N）	123	1157

由上表可以看到模型的精确率为0.925,召回率为0.888,F1-score为0.906。

将训练好的模型保存，将文本评论词向量化带入模型，进行预测，关于真实评论和虚假评论的情感预测概览，如下表所示。

表5.9 两类情感比重

	正向情感	负面情感	总计
真实评论	29681	3538	33219
虚假评论	11106	225	11331

由表5.9可以看到，在这两类评论内容当中，表达正面情感的评论都占据着大多数，负面情感只是评论的很小一部分。在模型预测的真实评论当中，正面情感的评论有29681条，负面评论有3539条，正面评论占比为89%，负面评论占比为11%；在虚假评论当中，正面评论有11106条，占比为98%，负面评论有225条，负面评论的占比非常低，为2%。由此我们可以看到，首先正面评论占比比较高，但是在虚假评论当中，正面评论占比尤其的高，达到了98%，可以得出，在这些虚假评论当中，充斥着大量的虚假好评，很可能是由商家雇佣的水军用户，进行好评操作。

5.6 行为特征分析

本节之前都是从文本字符内容的角度出发，对虚假评论进行分析和挖掘。但是，在文本内容之外，也可以分析虚假评论的行为特征。

5.6.1 会员等级

利用python爬虫获取到的数据可以知道每条评论的评论者是否为京东plus会员，基于评论者是否为plus会员的角度，对两类评论进行对比,如下表所示。

表5.10 会员比例

	虚假评论	真实评论
plus会员数	2386	18836
总数	11331	33219

比例	0.211	0.567

可以看到，在总共 11331 条虚假评论数据当中，有 2386 条为京东 plus 会员发布的评论，占据其总体的 21.1%；在总共 33219 条的真实评论数据当中，有 18836 条数据为京东 plus 会员发布的评价，占据总体的 56.7%。可以看到，在虚假评论当中，plus 会员的比例仅有 21.1%，而在真实评论当中，占据了 56.7%，显而易见，在虚假评论当中，大多数为非 plus 会员发布的评价信息，其评价质量比较低，而在真实评价当中，会员的比例比较高。

5.6.2 文本字符长度

对于评论文本的字符长度进行统计，字符长度均值为 83，字符最短为 4 个字符（删除 3 字符及以下），字符最长的评论为 576 个字符，可以看到大部分的文本字符长度都在 100 个字符以内。

对所有的文本数据按

照两类评论进行统计，然后观测其字符长度的分布如下所示：

表 5.11 虚假评论字符长度描述统计

	字符数		字符数
count	11331	25%	32
mean	69.70738	50%	51
std	57.00985	75%	92
min	4	max	535

表 5.12 真实评论字符长度描述统计

	字符		字符
count	33219	25%	51
mean	98.28853	50%	91
std	62.93383	75%	131
min	4	max	576

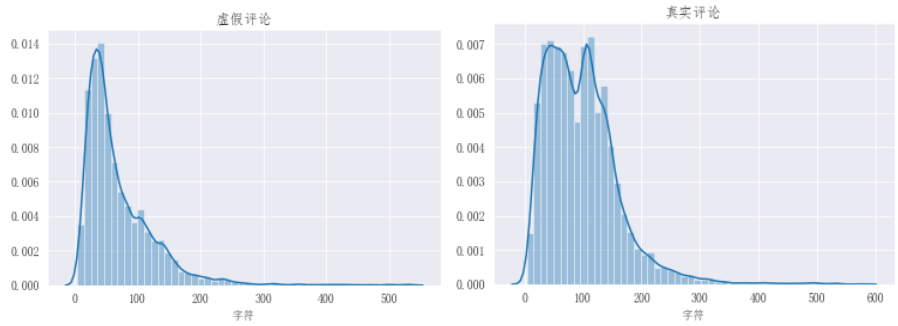


图 5.2 两类评论字符长度对比

对比两类评论可以看到，两者之间的分布有着明显的不同，真实评论的字符长度均值为 98.2，而虚假评论的字符长度均值为 69.7，有着非常明显的差异。相比较真实评论，虚假评论的字符在大体上要比真实评论要少。消费者往往会对商品会对所购买的商品简单的介绍它的优劣，和自己的使用体验，一般的消费者不会撰写特别长的评论，长度多控制在 100 字左右，然而虚假评论因为并没有什么切身的体验内容，所以其评论内容文本长度相对来说比较短。

6.结论

本文所研究的电商平台的在线评论数据属于非结构化数据的一部分，对于非结构化数据、自然语言的建模和分析是数据挖掘领域一项十分重要的工作。

本文首先通过 python 爬虫技术获取电商平台的某个品牌手机的在线评论数据，在通过一系列的数据清洗工作和文本向量化操作将原始文本数据处理成计算机可以理解的数值型数据，将向量化之后的文本通过传统的机器学习方法和深度神经网络方法建立基本的分类器，并通过优化对比保留最优的模型，保留五个基分类器（朴素贝叶斯、logistic Regression、SVM、LSTM、Text－CNN），最终采用集成学习当中的加权投票的方法对这些基分类器进行加权融合，形成最终的虚假评论识别模型，提升模型的稳定性和准确性，实现自动地对文本评论内容识别、判断，利用该模型进行预测。

通过利用虚假评论识别的模型对大量未标注的的评论数据进行预测，将这些未知数据划分成为真实评论、虚假评论两类评论数据，通过对两类评论数据进行对比分析，进一步分析虚假评论的语言特征和行为特征。

对虚假评论的特征分析主要从两个方面进行展开，一个方面是语言特征，另一方面是行为特征。

首先是语言特征，即从评论语料的角度出发，对比两类评论的语言差异。通过建立两类评论的统计语言模型以及分析 LDA 主题差异、词性分析等角度出发，可以看到对于虚假评论它的评论语料主要是对产品进行一些比较宽泛的评价，或者是针对产品的一些周边服务进行评价，如对客服的服务态度、快递物流等角度进行评价，相反真实评论主要是围绕产品的细节部分展开评价，对于产品细节、使用感受等方面的描述比重比较高，给出针对性的评价。从评论的词性角度进行分析，可以看到无论是哪两类评论，名词的比重都是最高的，在真实评论当中，形容词的比重要明显高于虚假评论，包含了很多对于商品具象的描述。虚假评论当中，对时间、数量大小等词汇的比例较低，较多地使用动词和副词。另外，本文对识别出来的两类评论进行情感建模分析，发现对于虚假评论来说，正面的情感评论占据着绝大多数，评论当中充斥着大量的虚假好评。

其次是行为特征，也就是从文本数据以外的角度出发，去发现两类评论的主要特征，对于两类评论的会员比例分析后发现，在虚假评论当中，plus 会员的比

例非常低，仅有所有评价数的21%，相反在真实评论当中，真实会员的比例为56%，会员的比例远高于虚假评论。另外，在对两类文本的字符长度进行统计后，可以看到两者之间的分布有着明显的不同，相比较真实评论，虚假评论的字符在大体上要比真实评论要少。消费者往往会对商品进行多角度的评价，也会较为明确的描述出商品的优劣，而相反虚假评论因为并没有涉及到实际的使用体验和感受，或是从完成任务节约成本，不愿意投入过多的时间编写评论的角度出发，字数往往会比较少。

本文通过文本挖掘的方式对虚假评论进行识别，随后利用lda主题聚类、词性分析以及统计语言模型等方法对虚假评论和真实评论的特点进行对比分析，分析虚假评论的语言特征和行为特征，尤其是针对真实评论里面消费者所表达的满意点和不满的地方，可以为商家改进自身服务、提升商品质量提供帮助。

另外，通过大量的在线评论的识别操作，可以统计出某个平台的在线评论质量状况，为整个平台的商品质量、用户活跃状况提供参考，帮助平台净化环境。此外还来可以根据虚假评论的比例判断平台的信用状况。

另一方面，不仅仅是平台和商家的层面。对于消费者来说，虚假评论的标注，能够帮助消费者清晰的识别虚假评论，客观的了解卖家的信用状况，更能够让在线评论发挥它应有的价值和作用，为消费者购物提供参考和意见，帮助消费者快速、准确地做出选择。

关于本文所研究内容仍有以下几点展望，可以在未来进行进一步提升：

1、本文旨在构造一套虚假评论识别方法，能够徐速、准确地对虚假评论进行识别，同时能够发掘出虚假评论的语言特征和行为特征，但是本文只是在单个平台选取了某个品牌的几家店铺的在线评论进行研究，模型的迁移能力可能不是很好，模型可能并不能兼容其他领域的商品评论，准确率会比较低。如果需要对其他领域的评论信息进行识别，需要对模型的训练数据进行增加，涵盖各个领域的数据，提高模型的泛化能力。

2、本文的研究因为考虑到成本的问题，标注的数据量比较少，所以模型的训练数据有限，可以通过扩充训练数据的方式进一步提高模型的准确率。另外，在线评论的没日增量都是巨大的，模型不能保持一成不变，需要不断的增加新的训练数据进行迭代，保持模型的有用性。

3、本文通过使用网络爬虫的技术来获取平台数据进行建模分析，能够获取的数据维度比较有限，有关于用户特征的数据更是少之又少，有关于用户的数据属于平台的隐私数据，外部用户并不能获取得到，如果能够获取到更加详细的评论用户数据，对于刻画用户特征将会十分有益，能够挖掘出更多的用户行为特征，丰富研究结果。

参考文献

[1]普华有策 ph：电子商务行业行业发展概况及趋势分析（附报告目录）[EB/OL]。https://baijiahao.baidu.com/s?id=1682473558568148052&wfr=spider&for=pc.

[2]Li H, Chen Z, Liu B, et al. Spotting Fake Reviews via Collective Positive-Unlabeled Learning[C]. IEEE International Conference on Data Mining. IEEE, 2015:899-904.

[3]Jindal N, Liu B. Opinion spam and analysis[C]. International Conference on Web Search and Data Mining. ACM, 2008:219-230.

[4]C. L. Lai, K. Q. Xu and R. Y. K. Lau, et al. Toward a language modeling approach for consumerreview spam detection. Proceedings of the IEEE international conference on e-Business engineering.2010: 1-8.

[5]Myle Ott, Yelin Choi, Claire Caridie and Jeffrey T. Hancock. Finding deceptive opinion spam by any stretch of the imagination[C]. In: Proceedings of the 49th Annual Meeting of the Association for Computation Linguistics: Human Language Technologies. 2011: 309-319.

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

[7]Yafeng Ren,Donghong Ji. Neural networks for deceptive opinion spam detection: An empirical study[J]. Information Sciences,2017,385-386.

[8] Lau R Y K, Liao S Y, Kwok C W, et al. Text mining and probabilistic language modeling for online review spam detection[J]. Acm Transactions on Management Information Systems. 2012, 2(4):1-30.

[9]Ee-Peng Lim,Viet-An Nguyen,Nitin Jindal,Bing Liu,Hady Wirawan Lauw. Detecting product review spammers using rating behaviors[P]. Information and knowledge management,2010.

[10]王曼,曹倩,孙践知,张青川,徐菲.一种基于集成学习的用户基础属性预测方法[J].小型微型计算机系统,2020,41(12):2509-2515.

[11]周超然,赵建平,马太,周欣.基于注意力机制和集成学习的网页黑名单判别方法[J/OL]. 计算机应用:1-7[2021-03-27].http://kns.cnki.net/kcms/detail/51.1307.TP.20201112.1134.004.html.

[12]杨云云. 电影虚假评论识别研究[D].苏州大学,2020.

[13]汪浩. 基于集成学习的虚假评论检测[D].沈阳理工大学,2020.

[14]刘秀. 基于神经网络的虚假评论识别系统的研究与实现[D].北京邮电大学,2019.

[15]杜茂康,叶琪.基于 PCA 与协同训练算法的虚假评论识别研究[J].计算机仿真,2019,36(02):452-457.

[16]张文宇. 基于行为分析的电子商务虚假评论者检测[D].云南大学,2018.

[17]道如那. 基于文本与用户行为挖掘的虚假评论识别研究[D].内蒙古大学,2018.

[18]张胜男. 基于深度学习的虚假评论检测的研究与设计[D].重庆大学,2018.

[19]韩俣谈. 基于卷积神经网络的虚假评论识别模型[J].科学技术创新,2018(06):78-79.

[20]顾松敏. 基于主题模型的虚假评论人群组检测方法[D].沈阳理工大学,2018.

[21]李静. 基于卷积神经网络的虚假评论识别技术的研究[D].北京邮电大学,2017.

[22]任亚峰,姬东鸿,尹兰.基于半监督学习算法的虚假评论识别研究[J].四川大学学报(工程科学版),2014,46(03):62-69.

[23]杨超,李天卓,谈森鹏,杨新凯.基于双卷积神经网络的虚假评论识别[J].计算机与数字工程,2020,48(08):1954-1957.

[24]孙晓燕,马路遥,乔娅丽. 基于文本特征融合的虚假评论识别[A]. 中国自动化学会过程控制专业委员会、中国自动化学会.第 31 届中国过程控制会议（CPCC 2020）摘要集[C].中国自动化学会过程控制专业委员会、中国自动化学会:中国自动化学会过程控制专业委员会,2020:1.

[25]朱宇航. 基于半监督学习的虚假评论检测方法研究[D].南京信息工程大学,2020.

[26]曾致远,卢晓勇,徐盛剑,陈木生.基于多层注意力机制深度学习模型的虚假评论检测[J].计算机应用与软件,2020,37(05):177-182.

[27]黄欣欣,年梅,胡创业,范祖奎.基于卷积神经网络的虚假评论检测[J].计算机时代,2019(11):41-45.

[28]荆云飞. 电子商务网站的虚假商品评论检测系统[D].上海交通大学,2019.

[29]王文琪. 虚假评论对消费者购买意愿的影响研究[D].北京邮电大学,2019.

[30]印佳明. 图书虚假评论的识别方法研究[D].北方工业大学,2019.

[31]贾少华. 基于LDA与PW-Word2vec的虚假评论识别方法研究[D].内蒙古大学,2019.

[32]吕海. 虚假产品评论在线检测技术研究[D].沈阳理工大学,2019.

[33]李洋,董红斌.基于 CNN 和 BiLSTM 网络特征融合的文本情感分析[J].计算机应用,2018,38(11):3075-3080.

[34]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.

[35]庞博. 虚假评论识别及评论有用性分析[D].哈尔滨工业大学,2018.

[36]蒋兵兵. 基于logistic回归的虚假影评人识别研究[D].华中师范大学,2018.

[37]道如那. 基于文本与用户行为挖掘的虚假评论识别研究[D].内蒙古大学,2018.

[38]王烨. 基于深度学习的垃圾评论识别研究[D].吉林大学,2018.

[39]李存林. 电商虚假评论识别[D].广西民族大学,2018.

[40]魏伟. 基于商品评论的情感分析研究[D].中南民族大学,2018.

[41]王禹,张世平,戴德. 电商网站虚假网购评论识别[A]. 中国统计教育学会.2017年（第五届）全国大学生统计建模大赛获奖论文选[C].中国统计教育学会:中国统计教育学会,2017:43.

[42]殷亚博,杨文忠,杨慧婷,许超英.基于卷积神经网络和 KNN 的短文本分类算法研究[J].计算机工程,2018,44(07):193-198.

[43]张恒. 基于深度学习的虚假评论识别方法研究[D].哈尔滨工业大学,2017.

[44]侯惠敏. 基于 Web 挖掘的虚假评论识别与推荐算法研究[D].西安电子科技大学,2017.

[45]皮琪. 基于深度学习的虚假评论识别系统的设计与实现[D].北京邮电大学,2017.

[46]皮琪,王文杰,杨飞,赵耀.基于深度学习的虚假评论识别[J].网络新媒体技术,2016,5(06):30-33.

[47]赵军,王红.融合情感极性和逻辑回归的虚假评论检测方法[J].智能系统学报,2016,11(03):336-342.

[48]龚千健. 基于循环神经网络模型的文本分类[D].华中科技大学,2016.

[49]任亚峰,姬东鸿,张红斌,尹兰.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(03):639-648.

[50]孟美任,丁晟春.虚假商品评论信息发布者行为动机分析[J].情报科学,2013,31(10):100-104.

[51]张玉芳,彭时名,吕佳.基于文本分类 TFIDF 方法的改进与应用[J].计算机工程,2006(19):76-78.

[52]周茜,赵明生,扈旻.中文文本分类中的特征选择研究[J].中文信息学报,2004(03):17-23.

[53]代六玲,黄河燕,陈肇雄.中文文本分类中特征抽取方法的比较研究[J].中文信息学报,2004(01):26-32.

[54]王继成,潘金贵,张福炎.Web 文本挖掘技术研究[J].计算机研究与发展,2000(05):513-520.

后 记

光阴似箭，日月如梭，三年的硕士研究生求学即将结束。回想期间的学习和生活，面对培育我的母校，心中无限感慨。在硕士期间，经过学校、老师们的培养，我在专业素养与做人做事方面都有了显著的提升。我衷心感谢母校和恩师们对我的栽培。历时近八个月，我完成了硕士学位论文。在此，我发自肺腑的感谢我的导师——杨盛菁教授。从入学以来，杨老师对我的要求就非常严格，也经常与我交流，倾听我的想法并为我指明方向。硕士研究生期间，我跟随老师，参与了多个课题、项目以及教材编写等事务，对自己专业素养的巩固与提升起到了重要的作用。跟随老师工作的过程中，老师教给了我专业技能，也以身作则，让我学习到了做人、做事踏实、严谨的作风。杨老师在我撰写论文的过程中也给我提供了很大的帮助。从论文定题，到思路的设计，提出创新的思想，再到论文的修改，多次对我进行面对面的细致指导。在此，我再次由衷的感谢杨老师。我即将正式离开校园，踏入社会，开始工作，以新的身份开始新的生活，但我不会忘记在校期间学到的知识和道理，我会充分利用我的专业技能，良好的完成工作，并在工作过程中不断学习，不断提升自己，是自己不断进步，也会恪守自己在校期间的精神与原则，做一个弘扬正能量，对社会有贡献的人。