

ARTICLE TYPE

基于电商平台的女装在线评论数据挖掘及情感分析

杨雨潇*

西安交通大学，西安，中国

*Corresponding author. Email: 2201111610@xjtu.edu.cn

Abstract

本文对某电商平台女士服装在线评论数据进行建模，包括文本进行预处理、探索性分析-单变量探索和多变量联合探索、英文分词并用停用词过滤、基于 VADER 的情绪分析模型、基于 Logistic 和 Lightgbm 算法的预测模型，从多个方面对电商评论数据进行了分析，为电商平台提供客户画像及情绪分析，可供辅助营销决策。

Keywords: 电商平台，在线评论，数据探索性分析，Logistic，LightGBM

1. 引言

随着电子商务行业的快速发展，网络购物变得越来越流行，销售模式由线下向线上逐渐转变，传统的营销模式面临改变。电商在线评论数据对消费者和商家都具有非常重要的价值，购买者通过各种电商平台平台购买商品，并对商品进行评论以表达对购物体验及商品性能的满意程度，同时，人们往往在购买商品之前，除了查看商品信息，也会浏览该商品的评论，从其他已购买的顾客评论中获取有价值的信息。因此，在人们无法看到具体实物时，评论成为了一个重要的参考渠道，影响着人们的购买决策。评论对商家也具有应用价值，顾客对商品的真实反馈可以帮助商家制定今后的商品改进或营销策略，而对在线评论进行文本情感分析可以有效提炼消费者体验信息和产品相应属性的评价观点，从而最直观地体现用户的需求和关注点，值得关注的是，在线评论数据不仅仅是评论文本内容，还包括评论的点赞数量、用户是否愿意推荐、用户的个人信息等等，获取这些信息有助于更深入理解消费者行为。由于女士更愿意在社交平台、电商平台上评论分享，并且其评论内容注重个体感受，因此女士服装是良好的研究领域。本文基于某匿名电商平台的女士服装在线评论数据，试图通过数据探索性分析、情绪分析、预测分析得到一些具有业务价值的信息，为电商平台提供相应的对策与建议。

2. 研究背景

评论文本数据与普通文本数据不同，评论不仅包含主题，还包含主观性的情感色彩，书写形式也不规范，具有口语化的特点，如何在挖掘主题的同时，进行情感分析成为一

个难题。情感是人们对商品的态度，是人们通过文本表达出的情绪，一般分为两种积极或消极，积极说明对这一属性较为满意，消极则是对商品的某一方面不满，没有达到预期。学者们认为用户评论具有高度的分析价值，虽然其研究角度各有不同，但都取得了一定的研究成果。在已有的研究中，一些学者主要从特征提取、情感分析等方面进行了研究。目前主流的文本挖掘模型有 LDA（主题生成模型）、最大熵模型，对于关键词提取方法，有 tf-idf、bm25、textrank、pagerank、互信息和左右熵等，对于词法分析，有 HMM、CRF、词性标注、命名实体识别等，对于句法，有句法结构分析和依存句法分析，对于文本向量化，主流方法有 tf-idf、word2vec、doc2vec、cw2vec 等等。

3. 样本数据

3.1 数据来源

本数据源自 2016 年某海外电商平台，为真实数据，由 nicapotato 爬虫爬取并经过脱敏化处理，整理后发布在 Kaggle 平台上。

3.2 数据描述

3.2.1 元数据描述

该数据集包括 23486 行和 10 个特征变量。每一行为一个用户的评论，每一列为一个特征变量，详细说明见表 1

表 1. 元数据

特征名称	特征描述	特征类型
Clothing ID	服装标签	整数类别变量
Age	评论者年龄	正整数变量
Title	评论标题	字符串变量
Review Text	评论内容	字符串变量
Rating	产品评价打分	正序整数变量，从 1 (最差) 到 5 (最佳)
Recommended IND	评论者是否推荐某服装	正整数 0-1 变量，其中 1 为推荐，0 为不推荐
Positive Feedback Count	该评论得到其他用户积极认可的数量 (评论点赞)	正整数变量
Division Name	产品类别	字符串变量
Department Name	服饰类别	字符串变量
Class Name	服饰类型	字符串变量

3.2.2 数据统计性描述

- 数值变量：

通过对数值变量的简单统计性描述（表 2），对于 Age-年龄，我们发现评论者年龄平均为 43 岁左右；对于 Rating-产品评价打分，其平均在 4.2 分，满分为 5 分，分数较高，

通过观察四分位数，我们可以简单得出其分布为左偏分布；对于 Recommended IND-是否推荐，其平均分为 0.822，通过观察四分位数，我们也可以简单得出其分布为左偏分布，可能与 Rating-产品评价打分具有较强的正相关性；对于 Positive Feedback Count-评论点赞，其标准差较大，说明数据分布波动较大，同时其分布为右偏分布。

表 2. 数值型变量统计性描述

特征名称	平均值	标准差	最小值	1/4 四分位数	1/2 四分位数	3/4 四分位数	最大值
Clothing ID	918.118709	203.298980	0.0	861.0	936.0	1078.0	1205.0
Age	43.198544	12.279544	18.0	34.0	41.0	52.0	99.0
Rating	4.196032	1.110031	1.0	4.0	5.0	5.0	5.0
Recommended IND	0.822362	0.382216	0.0	1.0	1.0	1.0	1.0
Positive Feedback Count	2.535936	5.702202	0.0	0.0	1.0	3.0	122.0

- 字符型变量：

通过对字符型变量的简单统计性描述（表 3），对于 Title-评论标题，出现频次最高的数据为正向情绪，出现了 136 次；对于 Review Text-评论内容，其内容多为不重复，这也在预期之中，因为评论内容具有较长的长度，出现重复的可能性也小；对于剩下三个特征，我们分别可以得知其出现频次最高的数据的出现频次。分别为 13850、10468、6319 次。

表 3. 字符型变量统计性描述

特征名称	不重复数据数量	频次最高数据	频次
Title	13993	Love it!	136
Review Text	22634	Perfect fit and i've gotten so many compliment...	3
Division Name	3	General	13850
Department Name	6	Tops	10468
Class Name	20	Dresses	6319

3.3 数据预处理

- 缺失值处理：

由缺失值矩阵图和柱状图（图 1 和图 2）我们发现 Title-标题数据缺失较为严重，Review Text-评论内容也有一定程度的缺失。在 23486 条数据中缺损值分别为：Title-标题 3810 条，Review Text-评论内容 845 条，Division Name-产品类别、Class Name-服饰类别、Department Name-服饰类型各 14 条。

对于缺失值的处理，由于 Review Text-评论内容是分析的重点，无法通过其他方法补充，同时 Division Name-产品类别、Class Name-服饰类别、Department Name-服饰类型

仅凭服装的 ID 无法获取，因此将该四个特征的所有缺失值所在的行删去。而 Title-标题由于缺失数据量较大，占整体数据的百分之 16，不能将其简单删除，如若将其与 Review Text-评论内容结合，又无法探索 Title 本身的信息，因此在此处先对其进行保留不处理，删除后，共有 2966 个缺失值，共有 22628 行数据。

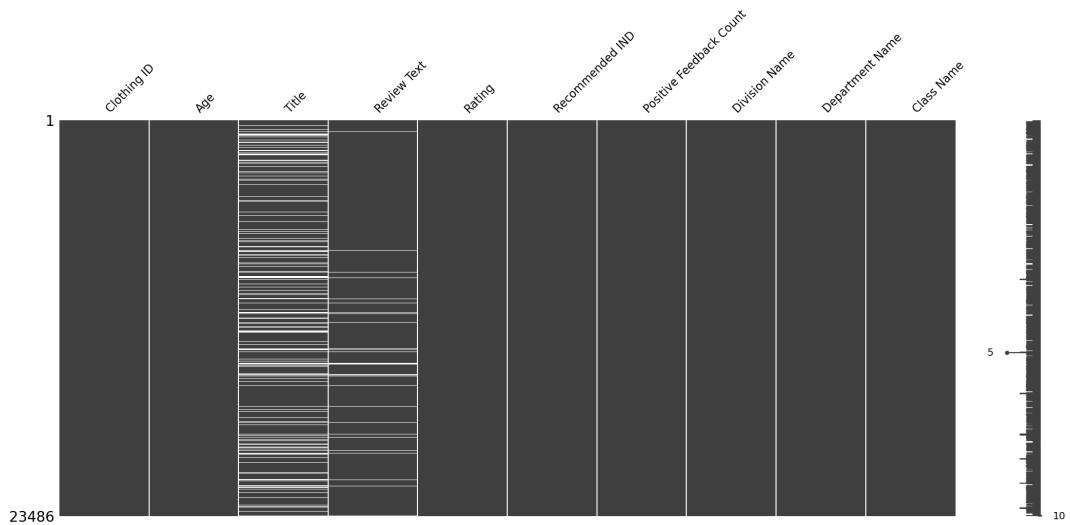


图 1. 数据缺失值分布

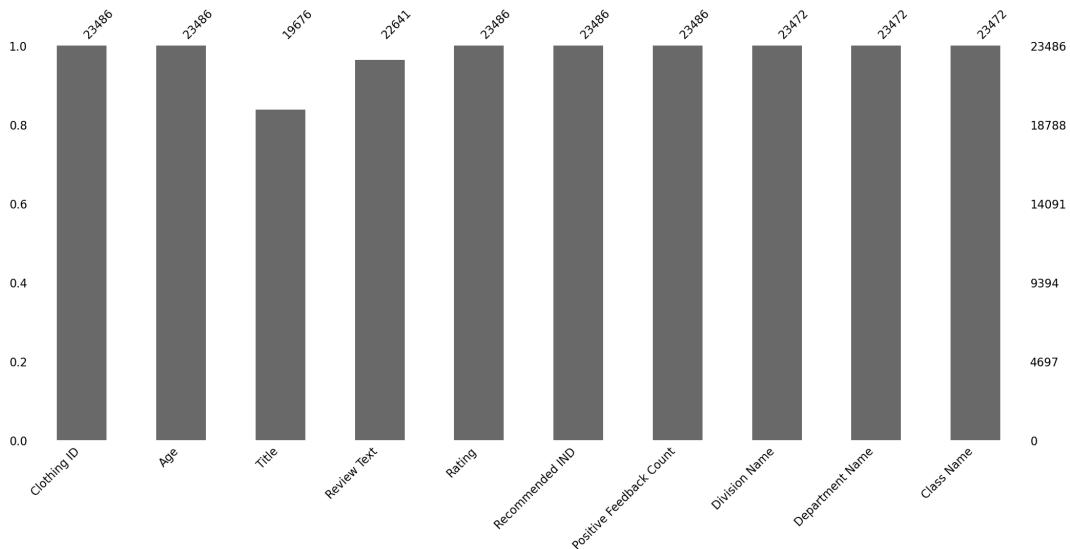


图 2. 缺失值柱状图

3.4 特征提取

观察特征变量，由于 Title-标题和 Review Text-评论内容为字符型变量，无法与其他数值型变量同时分析，因此可以添加描述字符型变量的特征变量，如对 Title-标题和 Review Text-评论内容的用词数量或者英文字母数量进行分析，因此在此基础上添加了两个新的特征，分别为 Review Text-评论内容的用词数量（Word Count）和英文字母数量（Character Count）；同时，为了便于之后进行自然语言的监督模型分类训练，我们需要将 Rating-产品评价打分二值化，即归纳为好和不好；值得一提的是，虽然 Recommended IND-是否推荐是二值 0-1 变量，但是该变量更能反映的是产品的社会性，即某个服装可能不适合评论者自身，但是它总体上是值得推荐给别人的，这就有悖于我们对产品评价打分的归纳；我们将大于等于 3 分的打分编码为 1，小于 3 分的打分编码为 0。

关于新提取特征的统计性描述分析：

表 4. 新提取特征统计行描述

特征名称	平均值	标准差	最小值	1/4 四分位数	1/2 四分位数	3/4 四分位数	最大值
Word Count	60.211950	28.533053	2.0	36.0	59.0	88.0	115.0
Character Count	308.761534	143.934126	9.0	186.0	302.0	459.0	508.0
Label	0.895263	0.306222	0.0	1.0	1.0	1.0	1.0

4. 数据探索

4.1 单变量数据探索

- Age-年龄和 Positive Feedback Count-评论点赞分布：

对年龄和评论点赞数量进行直方图可视化（图 3），可以发现电商评论中主要评论者为 40 岁左右的中年女性，呈现轻微的右偏分布，这与我们通常认为的电商消费评论者主要为年轻女性有些出入；同时，Positive Feedback Count-评论点赞数量好似呈现指数形式，大多数评论者都得到了极为少量的点赞，进一步研究，我对其先进行了平滑处理，之后取对数，探究其变化是否为指数，结果显示经过取对数后还有基本呈线性关系，这基本符合我们平常认为的分布，即少量的人占据着大量的资源，即二八定律，大多数都是低于平均值的。

对 Positive Feedback Count-评论点赞数量进行进一步的分析，自变量为当前评论者的累计百分比，排序由 Positive Feedback Count-评论点赞数量降序排序，因变量为当前评论点赞数量的累计百分比，得到的结果如图 4，我们可以发现，前百分之 20 的评论者拥有百分之 78 的点赞量，这几乎完美印证了二八定律，同时，我们计算 Positive Feedback Count-评论点赞数量变量的基尼系数，其值为 0.76，基尼系数越靠近 1，说明其不平均的程度越大；极少部分评论者的评价获得了大量的评论点赞，该部分评论者可能是最早期评论者，或者是其评论十分出彩，得到了大多数人的赞同而大多数人（约百分之 47）没有任何评论点赞。

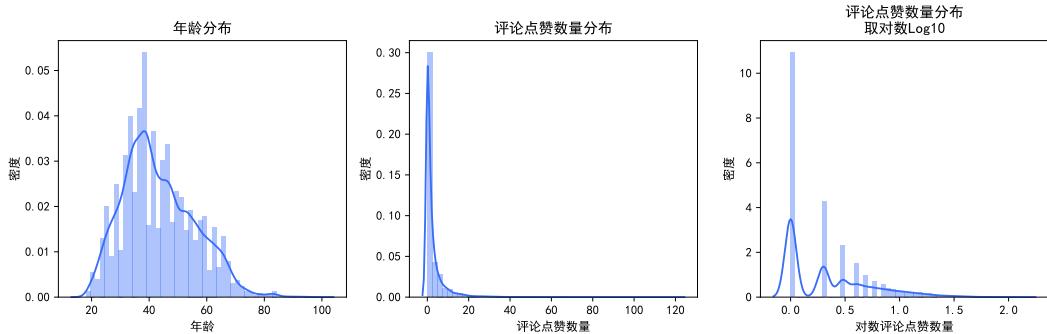


图 3. 年龄和评论点赞分布

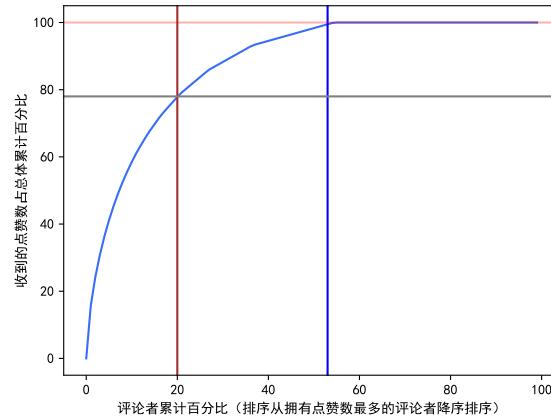


图 4. 评论点赞累计百分比图

•Rating-打分、Recommended IND-是否推荐和 Label-标签分布：

对评价打分、是否推荐和标签进行频数分布图可视化（图 5），通过 Rating-打分，可以发现该服装店的打分从 5 到 1 的数量逐渐减少，说明其店铺业绩表现很好，同时，通过是否推荐和 Label-标签和可以发现绝大部分的评分较高，客户的满意度较高，愿意将该商品分享给自己的亲友，分享到社会，说明该店铺的营销方案是相当成功的。

•Division Name-产品类别和 Department Name-服饰类别分布：

对产品类别和服饰类别进行频数分布图可视化（图 6），通过 Division Name-产品类别柱状图，可以发现购买普通款式的消费者最多，小款数量其次，亲密款消费者最少，这为我们构建用户画像提供了基础；通过 Department Name-服饰类别柱状图，可以发现上衣和连衣裙是评论数量最多的服饰类别，可以基于此研究这两种服饰类别的评论动机。对于 Class Name-服饰类型，由于其类型较多，因此放在多变量分布分析中进行。

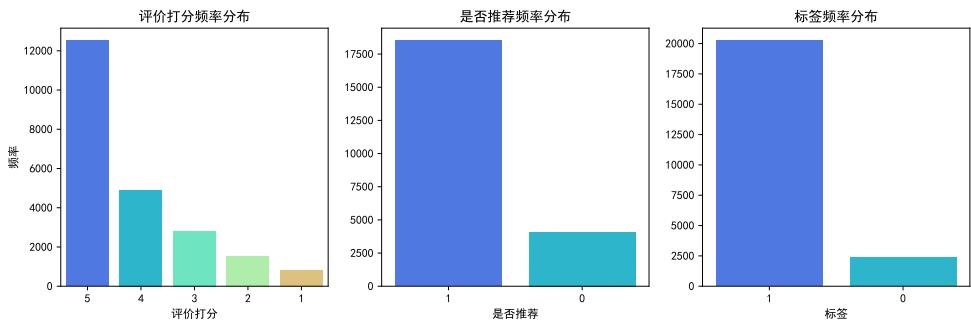


图 5. 打分、是否推荐和标签分布

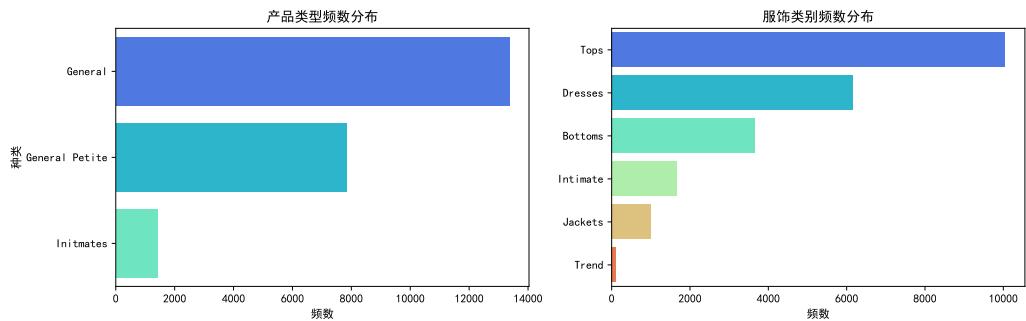


图 6. 产品类别和服饰类别分布

4.2 多变量数据联合探索

- 特征变量的相关性分析

计算数值型变量间的 pearson 相关性并以热力图可视化（图 7），可以发现 Label-标签与 Rating-打分具有强相关性，通过 Label-标签特征的生成方式我们也可以推出；同时，Rating-打分和 Recommended IND-是否推荐也具有强相关性，可以推出打分高的评论者更愿意将产品推荐给他人，而打分低的评论者更不愿意将产品推荐给他人；同时 Word Count-用词数量和 Character Count-英文字母数量与 Positive Feedback Count-评论点赞数量有弱相关性，可以推出评论内容的用词数量和评论点赞数量有弱正相关关系。值得关注的是，各特征之间没有强负相关关系。

- 基于不同 Class Name-服饰类型的平均年龄与是否推荐可能性的相关性分析：

经过研究各特征变量的相关性后，我选择了某个离散分类变量，研究在不同 Class Name-服饰类型的平均年龄与是否推荐可能性的相关性分析；计算不同服饰类型下平均年龄与是否推荐可能性的 pearson 相关性并以热力图可视化（图 8），可以发现推荐可能性-Recommended Likelihood 与平均年龄-Age Mean 有强负相关性关系，为了进一步研究关系，可视化了推荐可能性与平均年龄的散点图（图 9），可以发现随着年龄的增加，不同服饰类型的推荐可能性会降低；但这并不是因果关系，可能是因为某些服饰类型针

对的年龄段不同，同时其对于消费者的体验也不同，比如阿迪达斯 kids 针对的主要年龄段是童龄，其童装年龄打分也较高，而店内的一些其他物品（不针对儿童），受众年龄较大，同时由于不是主要产品，其打分可能较低。

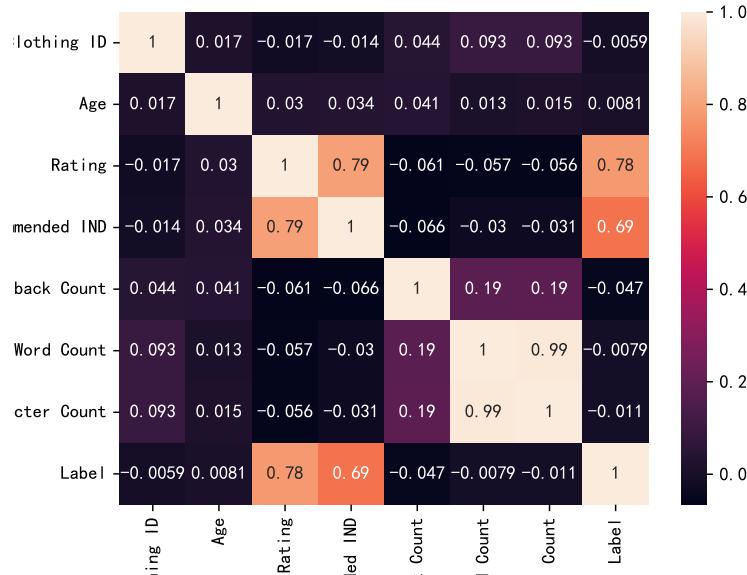


图 7. 各特征间的相关系数热力图

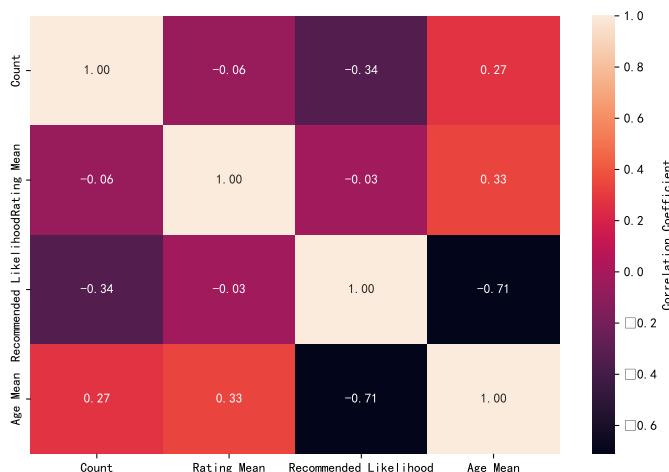


图 8. 基于不同服饰类型的相关系数热力图

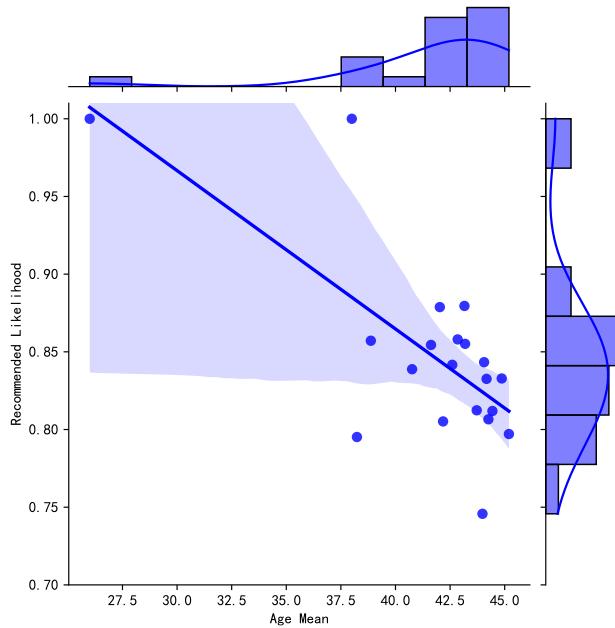


图 9. 平均年龄与推荐可能性散点图

- 基于不同 Rating-打分、Recommended IND-是否推荐、Division Name-服饰类别、Class Name-服饰类型的 Age-年龄频数分布

基于不同的特征变量对年龄进行频数可视化（图 10），在不同的特征取值下，年龄分布均无明显差异，说明打分、是否推荐、服饰类别、服饰类型等特征均与评论者年龄无相关关系，特征变量的变化不会影响年龄分布。

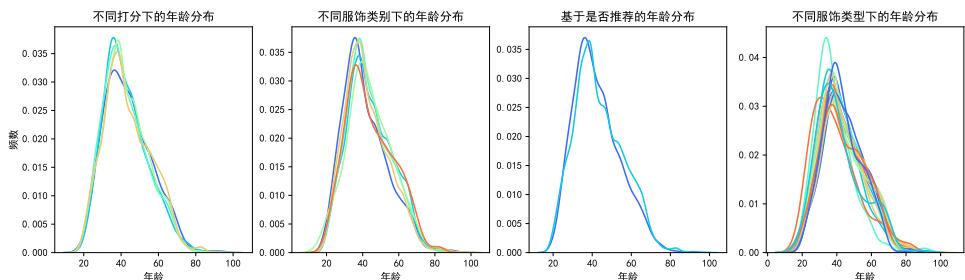


图 10. 基于不同特征取值的年龄分布

- Division Name-产品类别和 Class Name-服饰类型的交叉透视表

对产品类别和服饰类别进行频数与百分比分布热力图可视化（图 11），我们可以得出一些产品销量的信息，如销量最高的是普通款的连衣裙，普通款的针织物排列第二；在所有服饰类型中，销量最好的分别是上衣、连衣裙和编织衣物，且他们销量最高的都是普通款；连衣裙销量占整体销量百分之 27，其中普通款和小款分别占比 16 和 11。



图 11. 产品类别和服饰类型的交叉透视热力图

对产品类别和服饰类别进行两个维度的百分比分布热力图可视化（图 12），我们可以同样可以得出一些产品销量方面的信息，如 Chemises、Intimates、Layering、Legwear、Sleep、Swim 产品只有亲密款，Casual bottoms、Shorts 只有普通款，绝大部分产品普通款比小款销量占比大等。

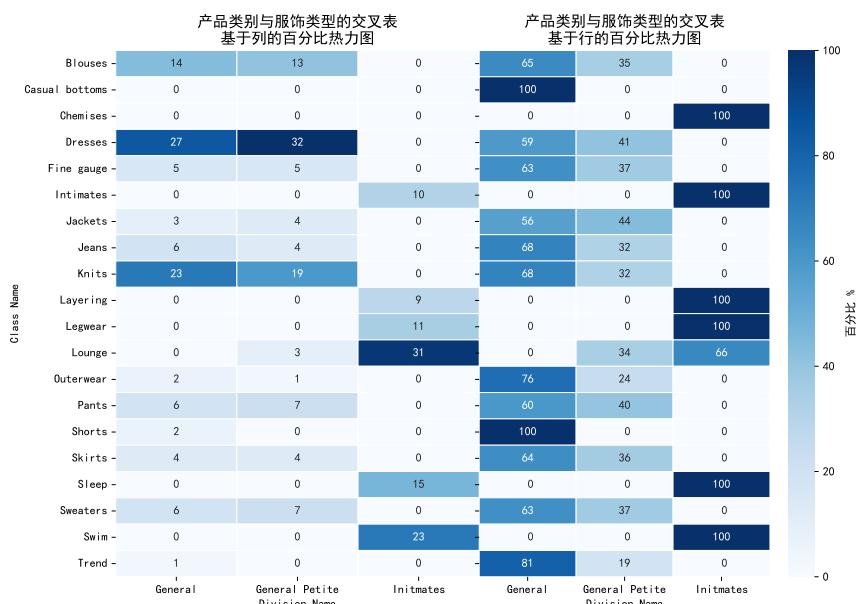


图 12. 产品类别和服饰类型的交叉透视热力图

5. 评论文本情感模型

5.1 评论文本处理

本文使用 nltk 自然语言处理工具包来进行评论文本的情感分析, 使用其中的 VADER Valence Aware Dictionary and sEntiment Reasoner 库来进行分析, Vader 于 2014 年开发, Vader(“价觉词典”和“情感推理者”)是一个经过预先训练的模型, 它使用基于规则的价值观来调整社交媒体的情绪。对于每一个文本, 给你一个评估, 不仅包括积极和消极, 以及这种情绪的强度。vader 在情感分析中会考虑否定表达(如, “not good”), 能表达情感信息和强度的标点符号(如, “Good!!!”), 情感强度(强度增强, 如“very”; 强度减弱如, “kind of”), 表达情感信息的俚语(如, ‘sux’), 能修饰俚语情感强度的词语(‘uber’、‘friggin’、‘kinda’), 表情符号:) and :D, 首字母缩略语(如, ‘lol’) 等等

使用 nltk.sentiment.vader 分析每一段评论内容的极性、正面、负面和中性分类, 其中, 参数 compound 表示复杂程度, neu 表示中性, neg 表示负面情绪, pos 表示正面情绪。将评论内容这四个维度存储到数据集中, 并且根据极性程度判断每条评论内容的情感倾向, 极性程度接近-1 时代表为消极情感极, 性程度接近 1 时代表为积极情感。

添加新特征后, 对这些特征进行探索性分析, 通过可视化情感倾向和评分的直方图(图 13), 发现具有积极情感倾向的文本占绝大多数, 而负面情感倾向的评论文本约占百分之十, 中立文本仅占极小部分。

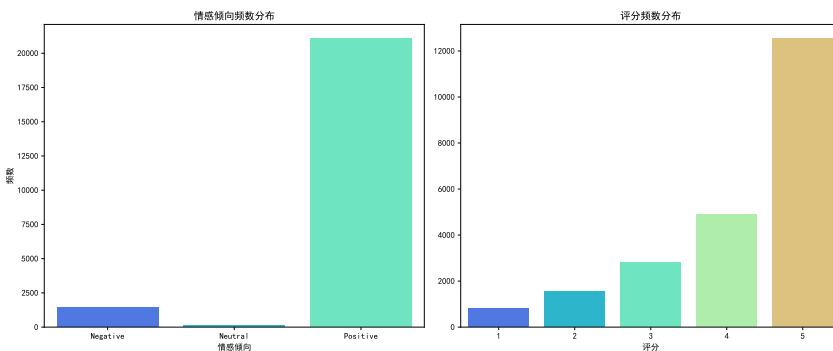


图 13. 情感倾向与评分对比图

5.2 词频可视化

利用词云的可视化功能, 单词出现频次与其大小成正比, 我们可以清晰地看到 Title-标题或 Review Text-评论内容中出现频次多的词汇。

图 14描绘了 Title-标题中出现频次多的词汇, 多为正向积极或中性的词语, 积极词语如 cute、favorite、perfect、please 等, 中性词语如 buy、design、major 等;

图 15描绘了高评价打分的 Review Text-评论内容中出现频次多的词汇, 同样多为正向积极的词语, 积极词语如 fit、love、pretty、comfortable、glad 等, 中性词语如 small、

store、layer、online 等；

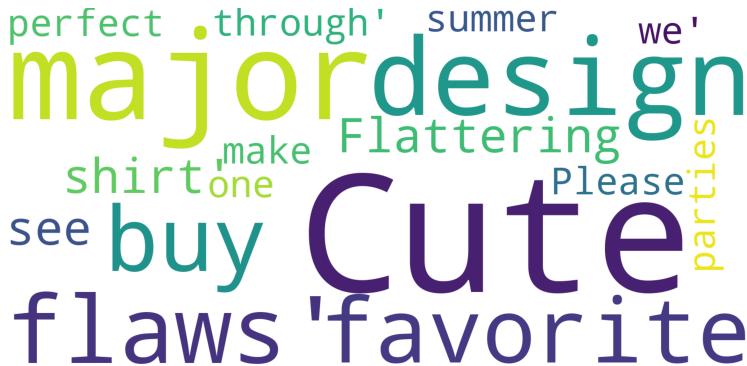


图 14. 标题词云图

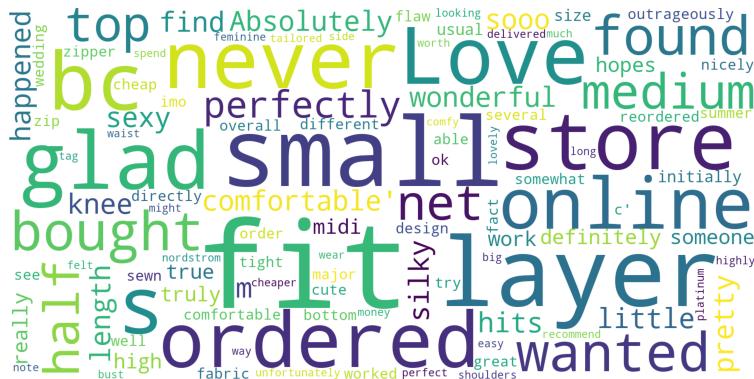


图 15. 高评分评论内容词云图

图 16 描绘了低评价打分的 Review Text-评论内容中出现频次多的词汇，出现了少量负面词汇，如 narrowing、sadly、waiting 等。

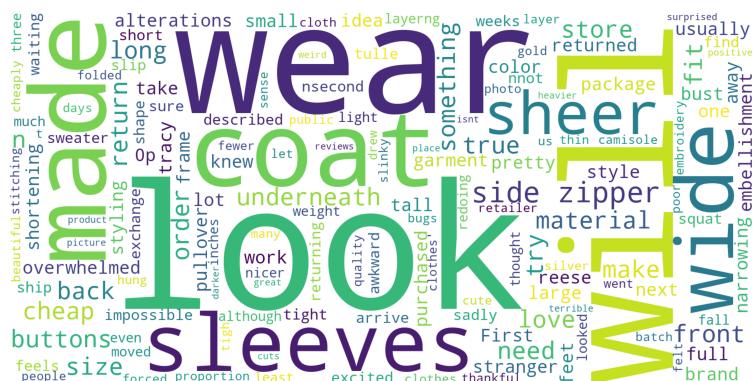


图 16. 低评分评论内容词云图

5.3 基于 Logistic 算法的模型训练与评估

- 模型训练

由于已经通过 nltk 自然语言处理工具包训练得到了评论内容的情感倾向，接下来我们想进一步探究评论内容与是否推荐之间的关系，即基于评论的内容训练出一个模型，推测出其是否有意愿将该产品推荐给好友，该模型将有助于我们判断在没有是否推荐属性时，人们的产品分享倾向，为之后的研究提供基础。

对于评论文本，利用 TF-IDF 来提取文本特征，使用 TfidfVectorizer 函数，将原始文本转换为 tf-idf 特征矩阵，将 CountVectorizer 和 TfidfTransformer 的所有功能组合在一个模型中；之后利用 sklearn 库的 train_test_split 函数划分训练集和验证集，最后利用 Logistic 算法进行训练，由于 logistic 算法中的参数对结果影响不大，在此使用默认参数训练。

- 模型评估

准确度：

Traindata Accuracy: 0.9128273118992376

Traindata ROC: 0.8887917239773266

Validationdata Accuracy: 0.8912947414935926

Validationdata ROC: 0.842036818826101

分类评估报告：

表5. 分类评估报告

attribute	precision	recall	f1-score	support
0(not Recommended)	0.56	0.78	0.65	596
1(Recommended)	0.96	0.91	0.94	3930
accuracy			0.89	4526
macro avg	0.76	0.84	0.79	4526
weighted avg	0.91	0.89	0.90	4526

特征权重可视化：（基于 eli5 库）

从图 17我们可以发现，love 为推荐标签最重要的特征向量，前三名特征向量都有积极的情感倾向；位于右列最底部的 disappointed 为影响不推荐标签权重最大的特征向量，但是仔细研究会发现，其中会有一些词汇如 was、wanted 等判断错误，说明模型还不是十分完备。

模型训练可视化分析

y=Recommended top features	
Weight?	Feature
+5.655	love
+5.176	perfect
+4.627	great
+4.527	little
+4.166	comfortable
+3.789	with
+3.490	soft
+3.319	compliments
+3.259	fits
+2.855	bit
+2.826	unique
+2.822	perfectly
+2.567	comfy
+2.563	happy
+2.522	jeans
+2.438	size
+2.391	glad
+2.365	slightly
+2.331	feminine
+2.311	nicely
+2.212	amazing
+2.135	easy
+2.133	fun
+2.125	bought
+2.099	gorgeous
+2.063	recommend
+2.056	flattering
+2.042	lovely
+2.007	casual
+1.986	justice
+1.950	saw
+1.947	beautifully
+1.937	order
-2.027	couldn't
-2.097	going
-2.118	disappointing
-2.159	however
-2.161	sack
-2.167	material
-2.168	idea
-2.208	nothing
-2.208	strange
-2.237	fabric
-2.245	shame
-2.263	hopes
-2.288	unfortunately
-2.320	completely
-2.408	thin
-2.450	awful
-2.474	were
-2.480	would
-2.500	return
-2.530	odd
-2.632	even
-2.674	maternity
-2.690	way
-2.775	excited
-2.987	poor
-3.125	bad
-3.196	back
-3.251	looked
-3.343	unflattering
-3.536	huge
-3.575	not
-4.048	returned
-4.059	was
-4.235	cheap
-4.311	returning
-4.560	wanted
-5.106	disappointed

图 17. 特征重要性可视化

Real Label: 0

y=Recommended (probability 0.757, score 1.135) top features

Contribution?	Feature
+1.072	<BIAS>
+0.062	Highlighted in text (sum)

the pattern is cute. it's a little more subtle than in the photos. the problems were the skirt are 1: pretty sure it will wrinkle in two seconds and 2: it has buttons down the front of it with no lining behind it. which means, if you sit down wearing this skirt and it doesn't lay quite right, it gaps and you can see whatever is or is not underneath! the buttons are like two inches apart, so it's not a tiny little gap. no thanks. also, this skirt fits true to size, it wasn't tight, just gappy.

Real Label: 1

y=Recommended (probability 0.938, score 2.709) top features

Contribution?	Feature
+1.637	Highlighted in text (sum)
+1.072	<BIAS>

it's rare to find clothing made of quality fabrics, and this is a fine, soft wool blend. the design is feminine without being over-the-top, and the fit is slightly loose, as pictured on the model, i would not recommend wearing it with a belt, as pictured; it makes it look too "trendy" and detracts from the shape of the cardigan, which is beautiful on its own just skimming the body. it is a perfect topper and can work for daytime or an evening out. really lovely, and one of the few things i decid

图 18. Logistic 算法下文本单词分析

5.4 基于 lightgbm 算法的模型训练与评估模型训练：

• 模型训练

LightGBM 是一款基于决策树算法的分布式梯度提升框架。为了满足工业界缩短模型计算时间的需求，LightGBM 的设计思路主要是两点：(1) 减小数据对内存的使用，保证单个机器在不牺牲速度的情况下，尽可能地用上更多的数据；(2) 减小通信的代价，提升多机并行时的效率，实现在计算上的线性加速。由此可见，LightGBM 的设计初衷就是提供一个快速高效、低内存占用、高准确度、支持并行和大规模数据处理的数据科学工具。

• 模型评估

train's auc: 1

valid's auc: 0.926843

F1 Score: 0.9365184109805992

模型训练可视化分析

Real Label: 1

'I love this shirt! it is so soft and a bit see through,
so recommend wearing a camisole underneath. perfect with
jeans! the fit is slightly boxy, but works well.'

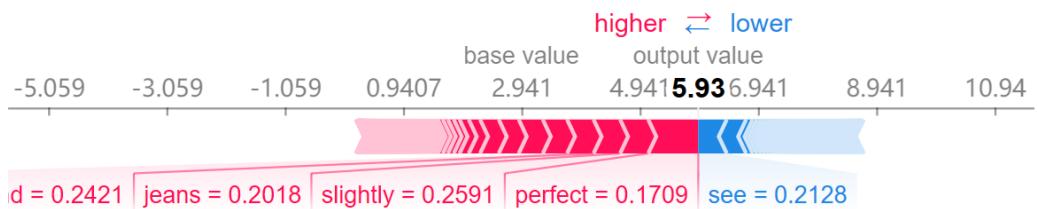


图 19. lightgbm 算法下文本单词分析 1

Real Label: 0

"This is way cuter on the model. it runs so big that it is very unflattering. perhaps on someone tall it would look better, but sadly it looked dumpy on my 5'2 frame. it also arrived with some black stain on it. it is going back."

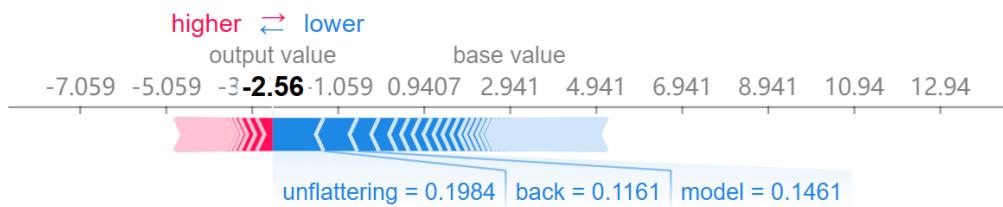


图 20. lightgbm 算法下文本单词分析 2

6. 结论与展望

本文针对某女装电商平台的在线评论数据进行建模，对在线评论数据进行预处理工作，包括缺失值处理、特征生成与选择；通过单变量和多变量分布的探索，对数据有了进一步的了解，同时发掘了一些客户画像信息；通过 VADER 的情绪分析模型实现了对文本评论数据的倾向性判断以及关于主题的高频特征词提取，并且通过高频词的频率分析，进一步提炼客户情感，从而为商家今后发展策略和其他客户购买产品提供进一步的建议。最后，通过 Logistic 和 Lightgbm 算法，将评论文本向量化，预测其推荐产品的概率，为商家设计新的宣传营销策略提供了启示。

7. 参考文献

- [1] 李杨, 徐泽水, 王新鑫. 基于在线评论的情感分析方法及应用 [J]. 控制与决策, 2023, 38(02):304-317. DOI:10.13195/j.kzyjc.2022.1788.
- [2] 杨春晓, 张鹤馨, 黄家雯, 等. 卷烟在线评论的文本情感分析 [J]. 中国烟草学报, 2020, 26(2):92-100.
- [3] 李宏媛, 陶然. 服装电商评论情感分析研究 [J]. 智能计算机与应用, 2017, 7(1):27-30,34. DOI:10.3969/j.issn.2095-2163.2017.01.007.
- [4] 胡云凤. 基于主题模型的在线评论分析方法研究 [D]. 陕西: 西安电子科技大学, 2017.