基于文本情感分析的商品评价模型

**Summary**

With the rapid development of the information age, the Internet and modern information technology are increasingly developed. The network has become the most convenient platform for data sharing, which also marks the arrival of a new era - the era of big data. In the era of big data, everyone is the creator and user of network data. Through big data, we can easily describe everyone's behavior habits and reflect the laws of things. E-commerce is the product of big data era. Founded in 1995 in Seattle, Washington, Amazon is the largest online e-commerce company in the United States. In the online marketplace it created, Amazon provides customers with an opportunity to rate and review purchases. In order to use these data to gain insights into the markets in which they participate, the timing of that participation, and the potential success of product design feature choices，it’s very meaningful to build a model to analyze user comments and ratings.

firsly, 为了帮助阳光公司提供的三个产品在亚马逊商城的销售取得成功，本文利用机器学习中的LightGBM算法建立模型, 通过自然语言情感分析技术将评论文字映射到数值空间, 同时对预处理之后的数据进行特征工程, 以帮助模型更好的分析数据, 建立模型后通过模型评价得到模型的F1值均在0.9左右, 表明模型建立的非常成功, 同时我们得出了数据中各指标的重要程度, 最终发现特征review\_body(评论), rate(好评率), review\_date(评论日期)和total\_votes(总投票数)对一个产品的评级影响很大, 故公司应注重相应的数据信息来成功运营产品.

secondly, 针对第二问的a小问, 在第一问时建立的模型就已经综合考虑了各种相关数据, 一旦阳光公司的产品开始销售, 根据反馈的评价信息, 将相关数据输入到模型中得到预测评分, 再根据预测评分和真实评分来分析数据得到相关结论. 对于b小问, 本文先通过熵权法确定了综合评级和评论的评价score结果来描述产品的声誉, 并通过月份销售量, 好评率, 差评率的统计数据绘制统计图进行基于时间模式的分析, 发现xxxxxxx; 由于前边两个模型都结合了相关的文本和评级的度量, 故c小问通过模糊综合评价方法利用问题一模型结论得出相关指标和b小问得出的score进行综合评判得到相应指标的权重, 利用权重计算综合得分, 评判一个产品的潜在成功与否; d小问要求分析特定的评级是否会引发某种类型的评论, 通过计算二者的斯皮尔曼相关系数和实际数据发现, 特定的评级会引发某种类型的评论; 最后, 通过构建特定描述符词典, 统计相关描述符的词频, 并利用热力图和词云可视化分析得出, “never”等否定词汇与一星级密切相关, 而”love”,”like”,”great”等积极词汇常出现在4星或5星评级中.

finally, 本文将以上数据挖掘和分析结果汇总成报告, 以书信的方式向阳光公司的市场总监提供, 真诚的希望本次结果能够为公司提供最好的帮助.

Keyword：

**Content**

1. **Introduction**

**1.1Problem Formulation**

随着信息技术的持续大爆发，数据量也正在呈爆炸性增长的趋势，如何在大数据时代充分利用数据提取出有用的信息，这是一个非常具有意义的课题。亚马逊公司作为美国最大的电子商务公司，对用户的行为数据十分重视，该公司在其网络购物商城中为消费者提供了评分和评论的机会，本文将利用这些数据建立模型提取信息，为阳光公司新推出的三个产品上市及售后保驾护航。

该题目设立的问题可以分为两个部分，其中第一部分仅包含一个问题，第二部分包涵6个问题，具体问题如下。

第一部分:

a.分析吹风机、微波炉、剃须刀这三种产品的数据集，定量地测定评定星级、评论和Helpfulness Rating之间的数学关系，帮助阳光公司的三个新的上市产品获取成功。

第二部分：

a.提供一种可以根据评级和评论数据提取信息的方法，当推出的三个产品在网络上销售后，阳光公司可以利用这种方法追踪产品销售情况。

b.分别讨论着三个产品的数据基于时间的度量和模式，根据数据随时间的变化，分析商品声誉的变化。

c.确定一种基于文本的度量方法以及基于星级的度量方法组合，探寻产品是成功还是失败。

d.探寻客户是否具有从众心理，客户的情绪是否会随着他人的评价而被调动。

e.判断文本的特征词与评级水平是否密切相关

f.向阳光公司的市场总监写一封信推荐我们团队的测评结果并讲清楚理由。

**1.2Model Goals**

Sunshine公司的吹风机、微波炉和剃须刀这三种产品即将上市。本文建立模型期望达到的目标是，合理利用数据提取信息，帮助sunshine公司能够根据产品的评论和评级追踪产品的销售和声誉的变化情况，以及判断产品评级和评论对消费者的影响。总体来说，我们团队建模的目的是为sunshine公司制定出正确的销售策略做出贡献。

1. **Assumptions**

为确保数学模型的建立顺利进行，现对模型做出以下假设：

1.假设评论是真实的，不存在刷好评和故意抹黑。

2.假设评论不存在反讽情况。

3.假设评论使用的都是常用单词，如good, recommend等

4.假设产品的质量保持不变

5.假设客户对同一种类的产品的观点一致，如用户对hair\_drye下的产品的观点相同的。

1. **Symbolic Explanation**
2. **Methodology**

**4.1GBDT算法**

GBD是一个集成多个决策树的机器学习模型. GBDT全称梯度下降树，在传统机器学习算法里面是对真实分布拟合的最好的几种算法之一,其具有有效性,准确性,可解释性, 在众多机器学习任务中取得了优越的结果. gbdt通过多轮迭代,每轮迭代产生一个弱分类器模型，每个分类器在上一轮分类器的残差基础上进行训练。对弱分类器的要求一般是足够简单，并且是低方差和高偏差的。因为训练的过程是通过降低偏差来不断提高最终分类器的精度。然而GBDT存在着致命的缺点—速度很慢, 需要对数据进行无数次的遍历.导致其应用有限.

**4.2模糊综合评价模型**

模糊综合评价方法是模糊数学中应用的比较广泛的一种方法。在对某一事务进行评价时常会遇到这样一类问题，由于评价事务是由多方面的因素所决定的，因而要对每一因素进行评价；在每一因素作出一个单独评语的基础上，如何考虑所有因素而作出一个综合评语，这就是一个综合评价问题。模糊综合评价是对受多种因素影响的事物做出全面评价的一种十分有效的多因素决策方法，其特点是评价结果不是绝对地肯定或否定，而是以一个模糊集合来表示。

1. **Modeling and Analysis**

通过一定的数据分析和处理, 利用机器学习中的lgb模型对数据进行建模, 最终得到定量的数学模式帮助阳光公司分析市场, 最终获得一定的成功.

**5.1Data Preprocessing**

The title presents three annexes, which respectively provide data on the three categories of hair dryer, dryer and pacifier: 11470, 1615 and 18939. Also, the three data sets provided contain product user ratings and reviews extracted from the Amazon Customer Reviews Dataset thru Amazon Simple Storage Service(Amazon S3). Meanwhile, in each attachment, each row of data contains information of 15 categories of the commodity, and each category represents the specific meaning as shown in the following table\*.

Table \*:

|  |  |
| --- | --- |
| marketplace (string) | 2 letter country code of the marketplace where the review was written. |
| customer\_id (string) | Random identifier that can be used to aggregate reviews written by a single author. |
| review\_id (string) | The unique ID of the review. |
| product\_id (string) | The unique Product ID the review pertains to. |
| product\_parent (string) | Random identifier that can be used to aggregate reviews for the same product. |
| product\_title (string) | Title of the product. |
| product\_category (string) | The major consumer category for the product. |
| star\_rating (int) | The 1-5 star rating of the review. |
| helpful\_votes (int) | Number of helpful votes. |
| total\_votes (int) | Number of total votes the review received. |
| vine (string) | Customers are invited to become Amazon Vine Voices based on the trust that they have earned in the Amazon community for writing accurate and insightful reviews. Amazon provides Amazon Vine members with free copies of products that have been submitted to the program by vendors. Amazon doesn't influence the opinions of Amazon Vine members, nor do they modify or edit reviews. |
| verified\_purchase (string) | A “Y” indicates Amazon verified that the person writing the review purchased the product at Amazon and didn't receive the product at a deep discount. |
| review\_headline (string) | The title of the review. |
| review\_body (string) | The review text. |
| review\_date (bigint) | The date the review was written. |

For a large number of commodity evaluation information data, in order to facilitate the subsequent establishment of mathematical model to solve the problem, the data is preprocessed according to the following steps:

Step 1: After checking all the data, it is found that only four rows of data have missing values, which has little impact on the overall data volume, so the four rows of data are selected for deletion.

Step 2:After careful observation of the data, it can be found that the values of all sample data in the two categories of market place and product category are the same, indicating that these two categories are irrelevant variables in this study, so they are deleted; at the same time, in the data in the column of product\_parent, except for the two outliers in the pacifier, all other data, as long as the product\_id is the same, The same is true for product\_parent, so only one field is reserved.

Step 3:Because product\_id and product\_title have the same meaning(都标志着一个产品) and data redundancy, delete product title.

Step 4:Add two columns of data, year and month, according to the data of review\_date.

Step 5: 对review\_body进行分词、去除停用词、去除标点符号等文本处理，方便后期对文本进行分析

**5.2LightGBM Model**

LightGBM是对GBDT的高效实现，加速了传统GBDT训练过程20倍以上，同时达到了几乎相同的精度。主要包括两个核心部分，一个是GOSS--在减少数据量和保证精度上平衡的算法。另一个是EFB--能够将许多互斥的特征变为低维稠密的特征,有效的避免不必要0值特征的计算。

GOSS算法排除了大部分具有小梯度的数据，只使用剩余的数据进行信息增益估计，LightGBM研究[1]表示:具有较大梯度的样本在计算信息增益的时候扮演着更加重要的角色，GOSS可以通过更加小规模的数据来获得非常精准的信息增益计算。

EFB算法通过将互斥的特征捆绑在一起，来减少特征数目。互斥特征意味着它们几乎很少同时出现在非零值，并且LightGBM也表明:找到最优化特征捆绑是NP问题，但是贪心算法能够获得非常好的近似概率

**5.2.1 Model Establishmen and analysis**

我们将按照下面的流程来构建模型解决第一问.



其中hair\_dryer, microwave, pacifier三个数据集为经过Data Preprocessing之后的数据集. 接下来详述以上流程

**5.2.2情感评分**

We use TextBlob model [加参考文献]which is used in Sentiment Analysis to quantify the reviews\_body. The result of TextBlob Sentiment Analysis can return a tuple in the form of(polarity, subjectivity).

|  |  |  |
| --- | --- | --- |
| reviews | polarity | subjectivity |
| Works great! | 1.0 | 0.75 |
| Love this dryer! | 0.625 | 0.6 |
| Quiet, but does not seem like 1000 watt power… | 0.0 | 0.333 |
| Amazing addition to the nursery! | 0.750 | 0.9 |
|  | ⋯ | ⋯ |

The Polarity is a floating point number with a range of [-1.0, 1.0]. Positive Numbers mean positive and negative Numbers mean negative.Alfred is a floating point number with a range of [0.0, 1.0], where 0.0 is objective and 1.0 is subjective. 我们用reviews计算出来的polarity值来代替原来的reviews\_body数据, 从而建立我们的模型.

当我们用TextBlob将review\_body映射到数值空间中后, We build Light gradient accelerator (Light) Gradient Boosting Machine (LightGBM) which is widely used in data mining tasks to get the weights to build model. We take star\_rating as the output and the other features after preprocessing above as the input to train the model.

**5.2.2特征工程**

In order to obtain more data correlation information, we used feature engineering to process the data set, and then put it into the model for training.Feature engineering often plays an important role in the field of data mining. Different feature engineering can obtain different feature sets. A good feature engineering can reveal more relevant information in the data set, thus making the model more accurate.

我们观察数据发现，有些评论和评级存在很大的差距性，例如给了好评，却只打了一星，给差评，却打了五星。

|  |  |  |
| --- | --- | --- |
| review\_id | Star\_rating | review\_body |
| R134FUK2D9TQU6 | 1 | I have used the dryer several times and it works great. I had questions which were answered promptly by other customers which was helpful in making my decision. Definitely recommend. |
| R1HI3QGXJQ2RUT | 5 | We owned these from the store and they are exactly the same. Too bad my grandson decided he was done with pacifiers one week later |
| … | … | … |

通过统计发现该类数据在数据集hair\_dryer, microwave, pacifier中分别存在12, 5, 10条,本文认为此类数据的出现为异常情况, 并非正常的评价信息, 故此后的处理中将该类数据进行删除处理.

同时, 我们根据一个product\_id的所有star\_rating信息进行统计, 利用如下公式计算出一个product\_id的好评率rate:

$$rate=\frac{\sum (star\_rating>=4)}{count\_i}$$

其中$count\_i$表示product\_id为i的数据总数

最后, 我们将数据中的各种字符串类别数据统一为从0开始的连续数字类别数据, 因为接下来的模型需要该形式的类别数据.

**5.2.2训练数据构造和模型训练**

所有数据经过特征工程处理之后, 训练一个二分类器. 故将评级在4星及以上的视为正例, 4星一下的视为负例. 并且将数据集的20%大小的数据量作为测试数据来评测模型,剩下的数据作为模型的训练数据训练模型.最终得到的数据量大小如下表:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Train data | Test data | total |
| hair\_dryer | 9163 | 2291 | 11454 |
| microwave | 1288 | 322 | 1610 |
| pacifier | 15130 | 3783 | 18913 |

硬件环境为16G内存+intel core-i78th, 软件平台为windows10. LightGBM模型相关参数如下: num\_leavers=64, learning\_rate=0.09, 然后输入训练数据训练模型.

**5.2.3 模型评价**

利用测试数据来评价模型, 评价指标为precision, recall,F1值，计算公式如下:

$$precision=\frac{TP}{TP+FP}$$

$$recall=\frac{TP}{TP+FN}$$

$$F1=\frac{2\*P\*R}{P+R}$$

其中, $TP$表示评价真实结果为正例并且模型预测的结果也为正例的总数,$FP$表示模型把负例预测为正例的总数, $FN$表示把正例预测为负例的总数.

测试结果如下表

|  |  |  |  |
| --- | --- | --- | --- |
| product | precision | recall | F1 |
| hair\_dryer | 0.812 | 0.982 | 0.889 |
| microwave | 0.846 | 0.903 | 0.874 |
| pacifier | 0.892 | 0.989 | 0.939 |

由上表可知：LightGBM模型在三种产品上的F1值均较高,证明模型的效果都很好.

**5.2.2结果分析**

上文成功建立了LightGBM模型, 并且我们利用模型可以得出下表数据:

|  |  |  |  |
| --- | --- | --- | --- |
| 特征名称 | hair\_dryer | microwave | pacifier |
| customer\_id | 0 | 0 | 0 |
| review\_id | 0 | 0 | 0 |
| product\_id | 0.0365 | 0 | 0.0095 |
| helpful\_votes | 0.0412 | 0.0389 | 0.0444 |
| total\_votes | **0.1380** | **0.1229** | **0.1396** |
| vine | 0 | 0 | 0.0047 |
| verified\_purchase | 0.0285 | 0.0286 | 0.0253 |
| review\_body | **0.2238** | **0.3299** | **0.2857** |
| review\_date | **0.2** | **0.2131** | **0.2015** |
| year | 0.0031 | 0 | 0 |
| month | 0.0539 | 0.0819 | 0.0555 |
| rate | **0.2746** | **0.1844** | **0.2333** |

表1.特征值权值

表中数值为各特征的相对权重, 我们可以看出, customer\_id, review\_id在三个数据集上的权重都为0, 说明二者对于阳光公司的市场分析并无帮助, 仅能作为标识符存在, 对结果无影响. 其次我们看到review\_body, rate, review\_date 和total\_votes的权重都相对较高, 证明他们对于一个产品的评级非常重要. 所以, 对于阳光公司而言, 要重点从以上四个方面入手来获得成功.

**5.3Task A**

一旦阳光公司产品销售, 利用第一题的模型进行预测用户评级. 若预测评级与用户实际评级相符且评级较好, 则通过第一问得出的重要性来分析产品的成功之处, 将成功的方面做到底, 保证产品口碑不下滑. 若评级较差, 同样根据重要性来分析产品在哪些方面做的不够, 接下来应该着重提升产品劣势方面. 实现逆转. 若与预测评级与用户实际评级不符, 则应分析这些用户是否为虚假用户或恶意用户, 从而有针对性的处理该类用户, 保证正常的销售市场.

**5.4Task B**

**5.4.1模型的建立**

针对提供的三个数据集, 本文利用之前的情感分析结果, 结合评级数据, 利用信息熵确定二者权重, 最后得到每个评价的得分, 利用该得分来分析时间模式.

信息熵有三个性质，单调性、非负性和累加性：

1.单调性, 发生概率越高的事件, 其携带的信息量越低;

2.非负性: 信息熵可以看做一种广度量, 非负性是一种合理的必然;

3.累加性: 即多随机事件同时发生存在的总不确定性的量度是可以表示为各事件不确定性的量度的和，这也是广度量的一种体现。

假设用户给商品评级和评价是两个事件和, 期望他们独立, 则有他们同时发生的概率为



由累加性可知:



满足两个变量乘积函数值等于两个变量函数值的和, 应该为对数函数, 再考虑到再考虑到概率都是小于等于1的，取对数之后小于0，考虑到信息熵的第二条性质，所以需要在前边加上负号。最后参考信息论之父克劳德·香农给出的信息熵定义有信息熵公式为:



将情感分析后的数据作为评价数据, 将star\_rating作为评级数据, 最终计算产品综合得分score如下公式



emotionScore为对review\_body情感分析的分值, starRating为用户的评级. a和b是利用信息熵公式计算二者的信息熵权值, 对三个数据集分别计算权值得出表xxx:

|  |  |  |
| --- | --- | --- |
|  | a | b |
| hair\_dryer | 0.1381 | 0.8619 |
| microwave | 0.0693 | 0.9307 |
| pacifier | 0.169 | 0.831 |

计算每个产品的score来分析每个数据集的基于时间的度量模式. 本文通过按月份统计每个月的整体销量, 好评率, 差评率等等数据进行可视化分析. 其中单月份整体销量, 好评率, 差评率计算公式分别如下:





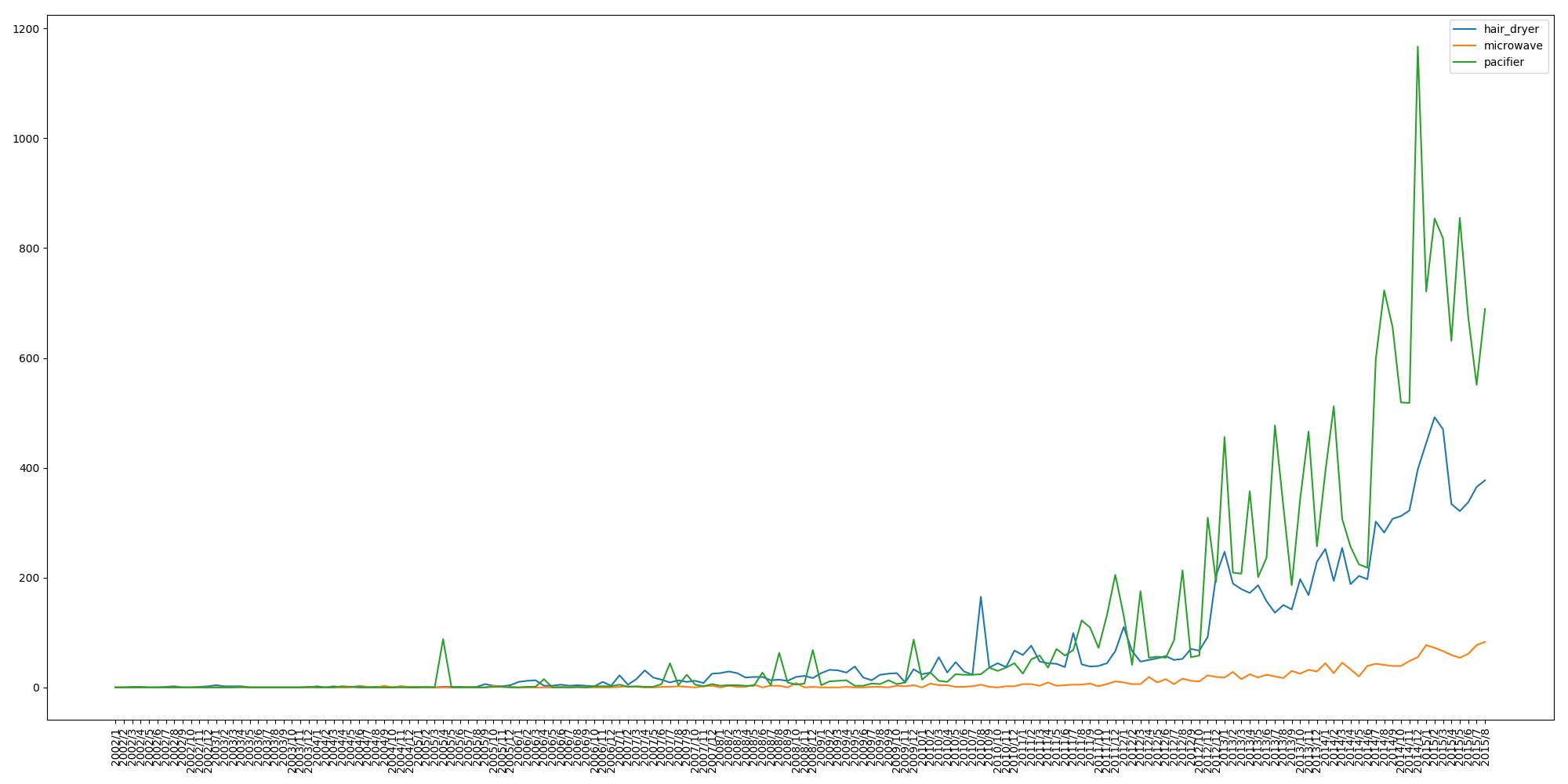


其中代表一行数据,即数据中的review\_date, 为真的每条数据利用公式计算的得分.

**5.4.2Result**

1.整体销量分析和数据选取

利用预处理的数据, 按月份统计三个产品的销量图如下xxx:

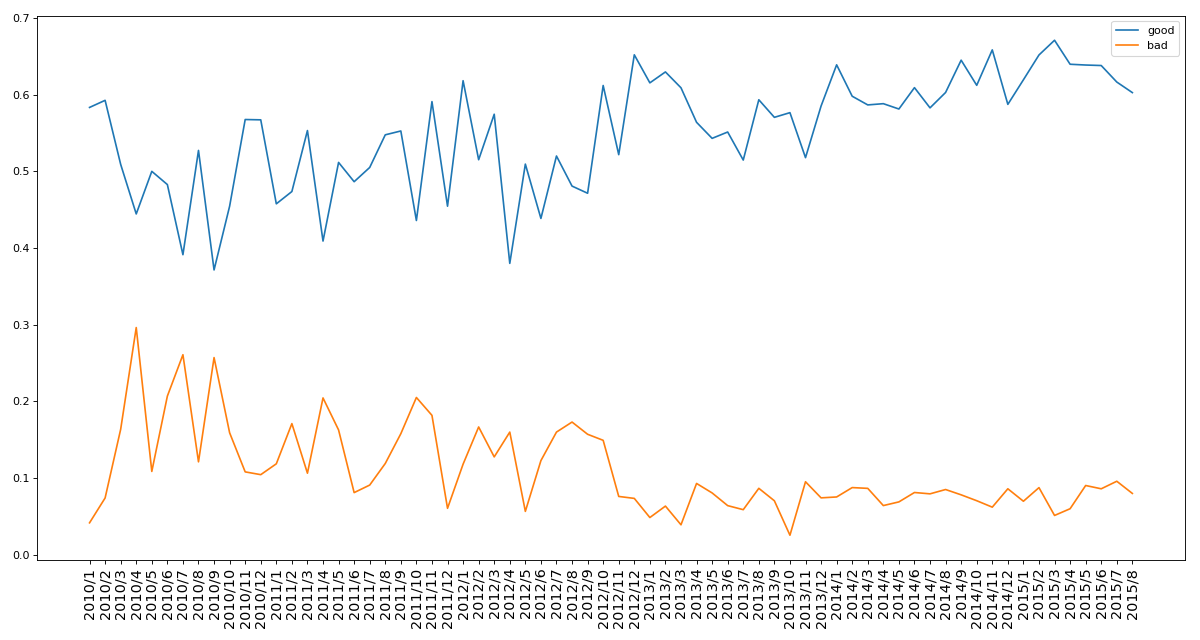


由于篇幅限制, 图像模糊, 清晰版请见附录, 通过图像可以得知2010年以前的销量非常少, 故下面分析时只取2010年1月以后的数据进行分析.

同时, 可以看出三个产品的总体销量在逐年上升, 微波炉销量相对较低, 尿不湿销量较高. 原因可能是微波炉属于耐用性产品, 而尿不湿则是一次性非耐用型产品, 吹风机销量居中. 这跟产品的特性和用途相关.

（1）hair\_dryer时间模式分析

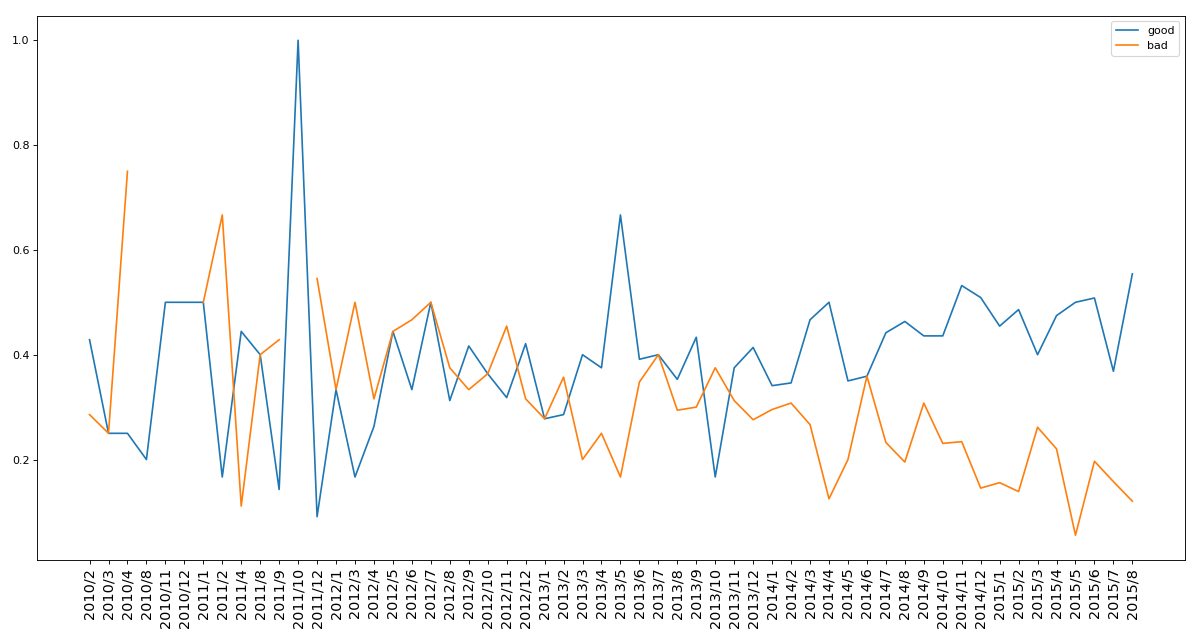
利用预处理之后的数据, 得到hair\_dryer的月份好评率和差评率如下图xxx:



从图中我们可以分析出从2010年1月以后, 吹风机的好评率总体上升, 差评率总体下降. 并且好评率始终远大于差评率. 并且在2013年1月份之前, 整个产品的差评率较高, 产品质量不客观, 但是2013年以后, 产品的好评率基本居高不下,说明吹风机在经历市场洗礼后逐渐质量口碑等逐渐上升, 人们使用起来更顺心. 但是仍会出现较轻微的波动.

（2）microwave时间模式分析

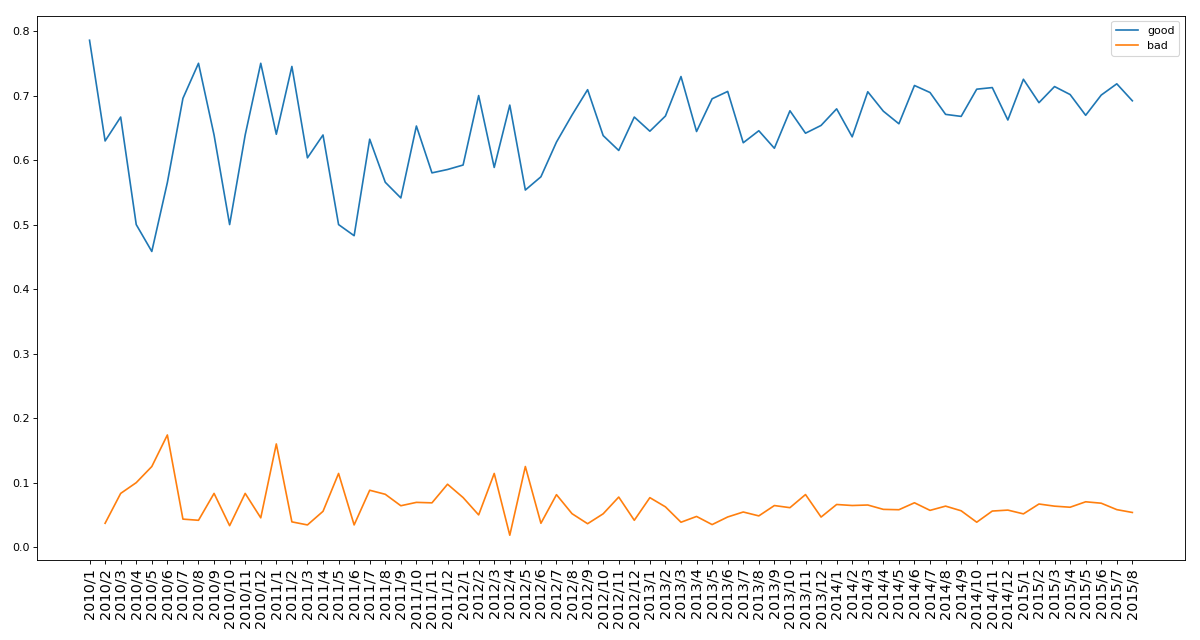
利用预处理之后的数据, 得到microwave的月份好评率和差评率如下图xxx:



从图中我们可以得出，微波炉投入市场的初期，市场对微波炉的差评率略高于高评率，微波炉的好评率再2011年10月达到峰值，随后下降；在2012年12月至2013年10月，微波炉的好评率和差评率在一定范围内波动并且较为稳定，在2013年11月和，该产品的好评率虽在一定范围内波动但恒大于差评率。总体来说，微波炉的差评率在下降，好评率在上升。

（3）pacifier时间模式分析

利用预处理之后的数据, 得到pacifier的月份好评率和差评率如下图xxx:



从图中可以分析得到，pacifier从产品投入市场到2015年8月，产品在市场中的好评率远远大于差评率，且波动频率越来越小，处于一种非常稳定的态势，是一个非常好的产品。

**5.5Task C**

要求确定一个基于文本的度量方法和基于评级的度量方法的组合, 来指示一个产品的成功或失败. 本文在第一题中建立了lgb模型并得出了相应指标对产品好坏的影响. 同时在上一小问中, 借助熵权法利用评级和评论的情感分值得到了产品的score, 上面两个解决方案都参考了评论和评级, 也即文本和评级, 故此处结合两个解决方案来最好的指示潜在的成功或失败的产品.

**5.5.1模型的建立与求解**

结合问题一的特征权重, 进一步筛选出有关文本的和有关评级的特征: review\_body, review\_date, rate, star\_rating. 同时结合上一小问的score. 建立模糊综合评价模型对产品的潜在成功与否进行评价. 模糊综合评价模型的基本步骤如下:

1. 选取特征
2. 确定评价矩阵
3. 检验评价矩阵, 计算权重
4. 计算得分, 得出结果

第一步选取特征已经完成. 接下来根据相关资料和数据确定评价矩阵A如下:



.通过对举着A进行一致性检验得出CR= 0.018<0.1 故矩阵通过一致性检验. 最后计算矩阵的特征向量并归一化后得到相应特征的权重如下:

|  |  |
| --- | --- |
| 特征 | 权重 |
| review\_body | 0.158 |
| review\_date | 0.098 |
| rate | 0.259 |
| star\_rating | 0.170 |
| score | 0.315 |

公司可以利用该权重值来评价产品的成功与否, 数值越大, 则产品越成功

**5.6Task D**

我们使用斯皮尔曼模型来测量评级与评论的相关性，斯皮尔曼模型计算最后的结果是一种秩相关系数，它是跟据原始数据的排序进行求解。

1.我们将每一条数据根据评论时间进行排序

2.将每一条数据的star\_rating和review的得分情况提取出来作为两个列向量R(Xi)和R(Yi)

3.将R(Xi)和R(Yi)输出到下面公式中：

d = \sum_{i=1}^{N} \left | R\left ( X_i\right )-R\left ( Y_i \right ) \right |^{2}

4.最后，根据下面公式计算出两个列向量之间的相关性Rs,得到结果如表1

Rs = 1-\frac{6 \times d}{N\times\left (N^2 -1 \right )}

|  |  |
| --- | --- |
| product | coefficient |
| hair\_dryer | 0.437746 |
| microwave | 0.559344 |
| pacifier | 0.372332 |

表1.系数

从表中可以得出，三种产品的相关系数大于0.05，所以，specific star ratings 会影响用户去编写某种类型的评论。并且我们在数据中发现部分评论确实会受到之前的评论的影响。并且我们在reviews中发现部分评分如表2中的评论。

|  |  |
| --- | --- |
| Review\_id | reviews |
| R3NQ2GXMX8FDFI | Here's a novel idea!! **Read the reviews BEFORE ..**. |
| |  |  | | --- | --- | | |  | | --- | | RK1L6FAK78AN4 | | | - DO NOT BUY - **the 5 star reviews are fake** |
| |  | | --- | | ROTACUL0CWK71 | | **I don't understand the great reviews for this dryer** |
| |  | | --- | | R3EPB70VVPJALX | | **Believe the Reviews!!!** |
| ⋯ | ⋯ |

表2 部分评论

从表中可以看出，评论中都提到了之前的评级评论等信息, 说明确实有些review会受到其他评级的影响。

**5.7Task E**

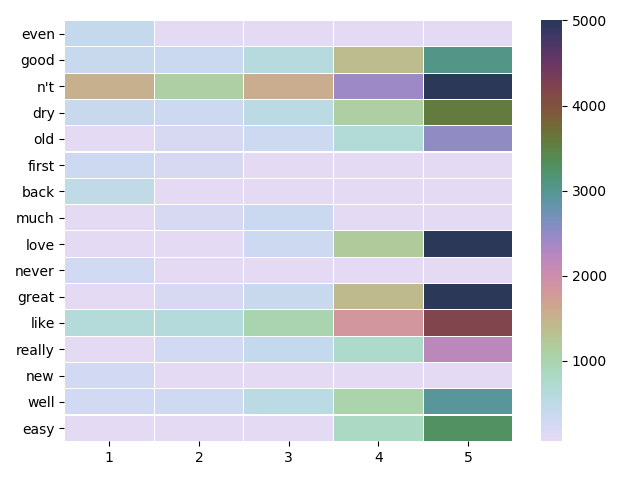
题目要求根据文本的评论的特定质量描述符, 和评级相关联, 挖掘其中是否有深层次信息. 为此本文通过对review\_body的具体内容进行挖掘, 提取其相关描述信息, 并通过统计和绘图等方式进行了挖掘.具体步骤如下:

1.建立特定的质量描述符词典. 通过查阅资料和相关数据, 本文建立了用户情感-评论-程度词典, 词典中包含英文的各种描述情感、评论词汇、程度词汇等等, 具体词汇表见附录.

2.将吹风机, 微波炉, 尿不湿数据合并, 同时提取review\_body内容. 按用户评级分类, 分别从review\_body中统计出现在词典中的词语, 通过词频统计方法得到各个词语出现的频次, 将结果按频次降序排列后取出前20个词语. 得到如下表格xxx:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 1 | 2 | 3 | 4 | 5 |
|  |  |  |  |  |  |
| old | 272 | 225 | 353 | 677 | 2512 |
| never | 298 | 0 | 0 | 0 | 0 |
| long | 241 | 186 | 327 | 563 | 1392 |
| new | 286 | 0 | 0 | 0 | 0 |
| good | 392 | 368 | 591 | 1402 | 3058 |
| much | 228 | 234 | 370 | 606 | 2079 |
| back | 460 | 212 | 0 | 0 | 0 |
| bad | 222 | 0 | 0 | 0 | 0 |
| high | 0 | 0 | 232 | 0 | 0 |
| however | 0 | 0 | 265 | 0 | 0 |
| need | 230 | 157 | 252 | 566 | 1645 |
| receive | 222 | 0 | 0 | 0 | 0 |
| keep | 0 | 196 | 261 | 479 | 1695 |
| cute | 0 | 0 | 0 | 0 | 1479 |
| perfect | 0 | 0 | 0 | 0 | 1600 |
| star | 0 | 0 | 0 | 440 | 0 |
| hot | 0 | 187 | 0 | 0 | 0 |
| easy | 0 | 0 | 0 | 832 | 3261 |
| dry | 386 | 326 | 535 | 1108 | 3578 |
| n't | 1509 | 1102 | 1594 | 2445 | 5932 |
| really | 237 | 268 | 417 | 784 | 2209 |
| want | 255 | 157 | 274 | 437 | 0 |
| love | 0 | 159 | 359 | 1184 | 8867 |
| great | 221 | 237 | 388 | 1405 | 5797 |
| nice | 0 | 0 | 0 | 558 | 1405 |
| small | 0 | 165 | 279 | 549 | 0 |
| like | 642 | 602 | 984 | 1851 | 4186 |
| best | 0 | 0 | 0 | 0 | 1448 |
| even | 401 | 208 | 233 | 0 | 0 |
| first | 355 | 223 | 0 | 0 | 1363 |
| still | 228 | 0 | 255 | 506 | 0 |
| set | 0 | 163 | 272 | 607 | 1378 |
| well | 285 | 302 | 531 | 1019 | 2932 |

利用表格数据绘制热力图(图xxx)进行直观分析.



图中横坐标为星级, 纵坐标为单词, 从图和表中可以看出, 评分为一星的评价出现的单词最多的是” n’t ”是个否定词, 出现了1509次, 接下来的是never, like等; 而评分为5星的出现最多的单词为love(8867次), 其次为great, like等等, 该单词在一星或二星中几乎未出现. 这表明特定的描述词汇与评级是密切相关联的. 同时, 我们也注意到了, “ n’t ”在各个星级中都出现了多次. 通过具体的review\_body数据分析, 在一星级中其多与其他积极词汇同时出现用以表达否定情感, 例如” n’t like ” 等等. 而在五星数据中其常形成许多积极情感的短语, 例如” like n’t allow finger print”等等.

通过上述分析, 我们总结出如下的星级常用词汇表xxx:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 1 | 2 | 3 | 4 | 5 | |
| not like | not good | not like | like | love |
| never | not like | not good | well | great |
| back | old | much | great | easy |
| even | not well | old | not | like |
| bad | dry | however | need | well |

* 1. **Task F**

Dear sir or madam:

非常荣幸有此次机会能够向您和贵公司展示我们对hair\_dryer, microwave, pacifier三个产品数据的挖掘和分析结果, 在这里我将向您全面报告我们的建模过程和我们在数据中的发现.

首先, 通过初步观测提供的三个数据集, 我们发现每个产品的marketplace都在US , 并且product\_category分别在每个数据集内部都一样, 我们认为其对分析无用, 故对其进行删除. 同样,发现product\_parent数据与product\_id数据冗余, 选择删除了product\_parent列数据. 同时, 对数据进行了其他预处理后, 进入建模阶段.

在经历自然语言处理中的情感分析过程后, 我们将非量化的评论文字映射到数值空间中, 用[-1,1]的数值范围来表示评论中蕴含的具体情感. 数值越大, 积极情感越浓厚, 数值越小, 则消极情感越浓厚. 情感分析完成后, 我们进行了特征工程, 一个好的特征工程能够让模型更好的抓住数据特征, 更好的发现数据中的内在模式, 提高模型的准确率. 通过特征工程建立了新的特征好评率rate和删除了一些非正常情况, 例如1星的评论:” I have used the dryer several times and it works great. I had questions which were answered promptly by other customers which was helpful in making my decision. Definitely recommend.”, 可是这应该是5星的好评才对, 所以针对此类评星和评级完全相反的数据进行删除处理, 幸运的是, 这样的数据并不多, 三个数据集一共只有27条. 随后, 我们将数据按8:2的比例进行划分训练集和测试集, 并将评级大于等于4星的作为正例, 小于四星的作为反例, 利用机器学习算法LightGBM对每个数据集进行单独的数据建模. 最终通过测试集的验证得到了模型在三个数据集上的F1值为0.889, 0.874, 0.939, 结果表明我们的模型建立的非常成功! 不仅如此, 我们还通过模型得知了review\_body, review\_date, rate 和total\_votes对模型的重要程度最大. 贵公司应该主要关注这四类数据的变化, 尽量少或不关注评论者等其他重要程度低的数据.

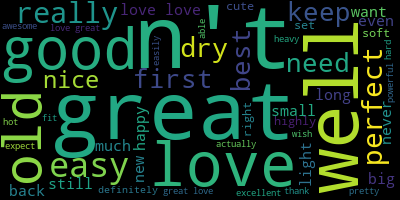
一旦贵公司打算上线销售三种产品, 可以利用上述模型对产品的销售数据进行评估, 通过模型预测结果和用户实际评分进行对比, 得到二者是否趋于一致的结果. 若一致, 则抓住此类用户的评价特点, 有针对性的进行改进产品, 以获得更好的销售. 若不一致, 则可以关注该类用户是否为虚假用户或恶意用户等等.

我们还利用了熵权法从评级和评价进行综合分析得到产品得分, 可以衡量产品的声誉等相关信息. 并且根据所给的数据分数据集讨论了基于时间的模式度量, 绘制统计图分析得出, xxxxx. 为了评价一个产品的潜在成功或失败, 我们结合了建立的机器学习模型LighGBM和熵权法分析的结果, 选取了重要程度大的与文本和评级有关的指标进行模糊综合评价建模, 最终得出可以通过计算特定分值来指示产品的潜在成功与否.

最后, 我们利用相关性分析技术得到了特定的评级会引发某种类型的评论, 即用户的评论可能会受到之前人评级的影响. 例如以下评论:” - DO NOT BUY - the 5 star reviews are fake”. 同时我们建立了特定词汇表, 对所有评论数据惊醒词频统计, 我们绘制了如下两张词云图, 词云图的单词的字体大小与其在评论中出现的频次有关. 频次越高, 字体越大. 通过词云图我们可以看到, 一星级的词云图大多为否定词汇, 例如n’t like, never, bad, 而五星级词云图则大多为great, love, well, perfect等. 虽然五星的词云图中也有n’t, 但是通过具体分析我们发现其主要表示积极情感, 例如” like n’t allow finger print”等等.



一星级词云图



五星级词云图

我们的数据挖掘结果呈现完毕了, 相信您看到之后一定会有大量的收获, 最后希望贵公司推出的产品能够大获成功!

1. **Conclusion**

通过对数据的具体分析

1. **Evaluating the Model**

我们将通过分析模型的优缺点和模型的应用来评价模型.

**7.1模型的优点**

1.我们对评论进行了量化处理，使用了TextBlob进行情感分析对每条评论进行打分，客观地得出了评论对产品的评级。

2.在解决问题d的时候，我们使用了斯皮尔曼系数分析，这种分析的要求不需要正态分布，并且它可以得出秩系数，即可以找出变量与变量以某种排序变化的关系。

3.模型较为客观的利用了尽可能多的数据, 并根据实际数据进行了调整, 通过对数据的可视化分析, 能够直观的挖掘出数据的大量信息, 透过信息来分析三个产品的趋势,声誉,成功与否等具体状况.

**7.2模型的缺点**

1.在构建LightGBM模型时, 由于时间问题未能很好的调参, 最终的结果虽然已经够好, 但是可能未达到最好.

2.模糊综合评价法的评价矩阵可能不够客观, 但是已经尽量通过数据保证了客观性.

3.词频统计时, 未能很好的将一些短语分析出来, 例如实际上想将”n’t like”等整体表现情感的短语分词, 但是未能很好的做到这一点, 导致最后的评级相关词中展现的不够完整

**7.3模型的应用**

1.模型可以应用在对产品初期的评估，分析其是否可以热销。

2.模型可以分析产品的声誉的变化，公司可以及时调整产品的销售策略。

3.模型可以较为客观的评价产品的成功与否, 并且及时向公司反馈相关信息.

1. **Reference**

[1][KE G, MENG Q, FINLEY T, et al. LightCBM: a highly efficient gradient boosting decision tree [C]// Proceedings of the 2017 Annual Conference on Neural Information Processing Systems. New York: Curran Associates Inc., 2017: 3146 -3154.]