

狮身人面之谜：破解亚马逊评级和评论的秘密

摘要

我们目睹了大规模在线市场的兴起。例如，最大的在线平台之一亚马逊的市值约为 9150 亿美元。在客户痴迷原则的指导下，为客户提供从 1 到 5 的打分机会。此外，购买者还可以提交一份基于文字的信息，即评论，来表达他们对产品的感受。这些评级和评论的海量数据提供了大量有待挖掘的信息。对基于文本的信息或基于评级的价值的分析已经得到了广泛的关注，但还没有一种方法可以将两者结合起来，特别是对于在线市场的情况。

为了解决上述挑战，我们提出了一种新的 CE-VADER 混合模型用于评论中的情绪分析，将消息分为强积极、弱积极、中等、弱消极和强消极五组。实证结果表明，所提出的五组分类模型与五星评价体系具有良好的相关性。在此基础上，提出了一种结合基于文本和基于评分的信息评估模型。我们挑选出 1%最有信息的评论和评级，以评估每个产品的属性，并提出销售策略。

我们在文献的微分方程模型的基础上，提出了“信誉”率来评价产品的信誉。然后采用自回归(AR)模型作为时间序列预测方法，预测未来的“信誉”率和每个产品的潜在成功或失败。AR 模型在验证集上具有较高的精度，最大均方根误差(RMSE)为 0.131。安抚奶嘴的名声很好，预计会成功，而微波炉和吹风机的名声不好，预计会失败。结果表明，连续五星或一星评级序列的比例具有相关性。最后，我们分析了具体的词汇和描述符，以找出它们与评分的相关性。

根据我们的实证结果，我们对网络市场提出了一些自信的销售策略和建议，如产品进入市场的时机选择，根据星级有针对性的调整等。我们给阳光公司的营销总监写了一封信，总结了我们的分析和结果，并提出了我们的建议。

我们的框架显示出了很强的准确性、鲁棒性。它可以很容易地用我们的源代码实现到其他数据。

关键词:基于文本的测量，信息性文本选择，声誉量化，销售策略形成



目录

1 介绍.....	3
2 假设和符号.....	3
2.1 假设.....	3
2.2 符号.....	4
3 信息化评价模型.....	4
3.1 星级的矢量编码形式.....	4
3.2 基于文本的度量的上下文熵 VADER 混合模型.....	5
3.3 基于文本和基于评级的措施相结合.....	8
3.4 模型实施、敏感性分析及结果.....	9
4 用差分方程测量基于时间的图案.....	10
4.1 基于差分方程的模型.....	10
4.2 模型实施，敏感性分析和结果.....	11
5 预测潜在的成功或失败.....	12
5.1 预测未来声誉的时间序列.....	12
5.2 评估成功或失败的可能性.....	12
5.3 模型实施与结果.....	12
6 具体评级和描述符分析.....	13
6.1 具体星级评级与评级频率的相关性.....	14
6.2 特定质量描述符与等级的相关性.....	16
7 设计特点的吸引力分析.....	17
8 销售策略和建议.....	17
9 优点和缺点.....	18
9.1 优势.....	18
9.2 缺点.....	18
10 结论.....	19
给阳光公司市场总监的一封信.....	19
参考文献.....	21
附录：.....	21



1 介绍

我们的社会见证了许多在线市场的崛起，其全球总市场价值已达 4.3 万亿美元。与传统平台相比，在线市场的一个显著特征是对文本的大量评论和评分。其中最受关注的是亚马逊，它最大的成功就是^[1]亚马逊还为顾客提供了自由表达自己感受的机会，并对他们购买的产品进行评分。

之前的工作^[2]表明，消费者在购买平台上的产品之前，很大程度上会参考评论和评级。平台可以通过查看这些评论来调整他们的销售策略。因此，评级和评论都为其他潜在买家提供了参考，也提供了大量的数据来分析客户的需求，有助于制定适应策略。充分利用这些数据，我们可以实现买家和平台的双赢。

其中最大的挑战之一是评论文本的复杂性和多样性^[3,4]。在本文中，我们提出了一种新的基于文本的情绪分析模型来解决这一问题。在本文中，我们开发了一系列模型，结合基于文本、基于评级和基于时间的措施，以挑选出最有信息的评级和评论进行跟踪。我们还建立了一个新的评估框架来量化每个产品的声誉和预测潜在的成功或失败。然后，我们分析了连续的相同星级评级、词汇描述符和产品声誉之间的相关性。我们在三种不同类型的产品，即奶嘴、微波炉和吹风机生成的真实数据集上实现我们的模型。

研究人员指出，有必要研究网络平台何时以及如何根据消费者评论或评分^[5]调整营销传播策略。基于我们的分析和结果，本文提出了几种销售策略和建议。

本文的其余部分组织如下。在第 2 节中，我们列出了模型构建中的主要假设，并介绍了将在本文中经常使用的符号。第三部分提出了一种新的信息评价模型。它是由一个混合的先进的 CE^[6]和 VADER^[7]在评论文本的情绪分析。然后，我们提出了“重要性”率作为一个组合的文本测量(即和基于评分的度量(例如，我们提出的 CE-VADER 模型)。(如星级评分和有帮助的投票)，以表明评论和评分的信息量有多大。据我们所知，我们是第一个提出基于评论文本的情绪分析模型的人。在第 4 节中，我们使用一个差分方程模型作为骨干来测量每个产品的时间模式。此外，本节还提出了“声誉”率来衡量声誉的增长或下降。在第 5 节中，我们使用一个自回归模型(AR)来预测未来时域的声誉变化，并提出一个模糊系统来预测每个产品的潜在成功或失败。关于我们在给定数据上实现的模型结果的更多细节可以在第 6,7,8 节中找到。第 9 节讨论了所提议的模型和框架的优缺点。我们在第 10 节结束。所有的源代码都附在附录 D-I 中，可以很容易地实现到其他数据集。

2 假设和符号

数学建模系列课程资料代码等请关注公众号：科研交流

2.1 假设

为了简化我们的模型并消除复杂性，我们在本文献中做了以下主要假设。一旦在我们的模型中使用了所有的假设，它们将被再次强调。

假设 1. 网上市场运行稳定。没有出现疫情爆发等严重影响网购产业链的情况。

假设 2. 评级和评论描述了顾客对他们购买的产品的真实体验和感觉。评语中的情愫，反映了一个人对产品的感受。



假设 3。顾客的经济地位、文化程度等绝大部分的个体差异被忽略了。

假设 4。运送产品需要一些时间。有些客户更喜欢在收到购买的产品后进行评论。

假设 5。消费者在购买产品时，更注重负面的评价，如低星级或负面评论。

2.2 符号

在这项工作中，我们使用表 1 中的命名法来构建模型。其他不常用的符号将在使用后引入。

表一 列出了本文中使用的符号。

Symbol	Definition	Type
id	review id	String
s_{id}	Star rate, subscript is its associated review id	Scalar
$h_{v_{id}}$	Helpful votes, subscript is its associated review id	Scalar
R_{id}	Review text, subscript is its associated review id	String
rd_{id}	Review date, subscript is its associated review id	Date
VEC	Vector encoding of the star rating	Mapping
INT	Vector encoding of intensity relevant to 5-class seed words	Mapping
IMP	Importance rate of review and associated rating	Mapping
REP	Reputation rate of product at some time	Mapping

3 信息化评价模型

在本节中，我们提出了评估评论文本和星级评价的信息量的“重要性”。我们考虑的最有信息的因素是复习文本的情绪。在这篇文献中，我们提出了一个 CE-VADER 模型来解决回顾文本中的情绪分析问题。我们的模型将文本分为五组：强、正、弱正、中、弱、强负面的五星评级方案一致性然后，我们提出的“重要性”将包含基于文本的度量，星级评级与他们的忠诚度，相关性。重要性越高，信息就越丰富。

本节的其余部分安排如下。在 3.1 节中，我们将积分星率转化为向量形式。在第 3.2 节中，我们提出了 CE-VADER，一种用于基于文本的度量的混合模型。在第 3.3 节中，我们介绍了“重要性”来计算评论和星级评价的信息量。在 3.4 节中，我们在 3 类产品的真实数据集上实现了我们的模型，以显示 1% 最具信息性的评论和星级评价，并分析了模型的敏感性。

3.1 星级的矢量编码形式

消费者购买后可以在亚马逊上自由表达对产品的评价，1 - 5 星。一星的满意度最低，五星的满意度最高。1 - 5 星的评级本身就足够了。将基于评级的度量与我们将在下一节中讨论的基于文本的度量结合起来。在本节中，我们将把星型评级转换为向量形式。

首先，我们根据给定的数据分别计算吹风机、婴儿奶嘴和微波各星级率的比例，如图 1 所示。我们观察到婴儿奶嘴获得最高百分比的高星级评价，而微波较低的星级评价。高技术含量的产品也会面临更多的质量问题，这是符合实际预期的，说明星级评级确实可以反映消费者的满意度。

然而，我们希望将评级转换为一个等效的 5 维向量编码形式。表示星率 $s \in \{1, 2, 3, 4, 5\}$ ， s 的向量编码形式可以表示为 $VEC(s) = (VEC)1s, \dots, vec5s)^T \in R^5$ 其中组件定义为：



$$vec_s^i = \frac{e^{-\frac{|i-s|^2}{2\sigma_0}}}{\sum_{j=1}^5 e^{-\frac{|j-s|^2}{2\sigma_0}}}$$

在 σ_0 是一个可调参数, 决定了模型的鲁棒性, 模型越大鲁棒性越强。VEC 映射是一一对应的, 因此我们声称转换的形式等价于星的评级。此外, 根据我们的定义, 我们可以得到: i) $s = \text{argmax}_i \{vec_s^i\}$; (二) $P5_i = 1 - vec_s^i$ 。VEC(s) 编码为概率向量, 每个分量表示其被相关星评级的可能性, 例如, 4 星评级最高的概率被评级为 4 星, 第二高的概率被评级为 3 星或 5 星。

3.2 基于文本的度量的上下文熵 VADER 混合模型

在这篇文献中, 我们在回顾文本的基础上构建了一个新的情感分析模型。为了简单起见, 情绪得分的文本被认为是唯一的事实来衡量产品的成功或失败, 积极的态度通常表明更高的潜力产品的成功, 而恰恰相反, 消极态度表明更高的可能性为产品失败。

在本节中, 我们提出了上下文熵和 VADER[7] 混合模型, 即 CE-VADER, 以解决审查文本中的情绪分析挑战。模型制作完成以上两个块: 上下文熵(CE) 块和维德块。CE 模型对股票市场新闻[6] 的情绪分析能力较强, 但也存在一定的局限性。



图 1: 根据给定数据, 奶嘴、微波炉、吹风机的星级评级分布。

背景例如, 文献综述。虽然 VADER 模型在简短的在线纹理方面优于先进的自然语言模型, 但其准确性很大程度上取决于预先列出的词汇。这两个 CE-VADER 块将分离出一个 5 维概率向量, 其中每个分量表示作为其关联组的概率。在一个投票群体之后, 我们提出的模型将把审查分成五组中的一组, 同时用一个强度来显示它被分成这一组的强度。通过杂交两种模型, 我们表明 CE-VADER 可以根据星级评级的一致性将评论上下文分为五组。

本节的其余部分安排如下。首先给出了生成 CE 模型种子词的策略, 并扩展了 VADER 块的金标准列表。然后对 CE 和 VADER 型号进行了比较详细的介绍。最后, 我们提出了 CE-VADER 混合模型。我们的模型将评价情境根据其强度分为强积极、弱积极、中等、弱消极和强消极五类。在下节中, 我们将提出代码基分类结果和强度的保真度审查。





图 2：我们提出的 CE-VADER 模型的整体架构。
该模型由两个模块组成，即 CE 模块和 VADER 模块。

3.2.1 手工注释种子词

我们把 80% 的数据作为训练集，剩下的 20% 作为评估的测试集。将训练集中的复习主体中的句子分解为独立的单词，统计出它们的出现频率。高频的情感词被我们挑选出来作为种子词手工标注，而低频的则被丢弃。注释者(我们小组的成员之一)将结合他的专业知识，自然语言处理知识，将所有选择的情感词分类为五组，即。、强正、弱正、中等、弱负、强负。我们没有对[6]进行“积极”或“消极”的粗糙的两组注释，而是详细地将每个注释细分为“强”和“弱”子组，并留出另一组标记为“适度”。五组注释策略的目的是与 1 - 5 的评分相关联，例如，“弱正”映射到四星评级。我们将种子字的五组表示为 G_i ，其中 $i = 1, 2, 3, 4, 5$ 。

从带有“正”或“负”标签的训练集生成的代表性词如图 3 所示，分别使用冷色调和暖色调。注释的五类种子词附在附录 A 中。



图 3：一些有代表性的种子词的演示。
注释为“积极”的词是浅色的，而“消极”的词是深色的。尺寸越大，词频越高。

3.2.2 上下文熵块(CE)

上下文熵块采用部分上下文熵模型^[6]作为主干架构。部分上下文熵模型既可以考虑共现的强度，也可以考虑来自审查上下文的最具代表性的候选词与生成的种子词之间的上下文分布。



我们使用一个矢量来编码单词和它的上下文之间的强度。更具体地说，表示 k 的左右上下文 th 词 w_k 在 n 字的复习上下文中， $R = w_1w_2 \cdots w_kw_{k+1} \cdots w_n \{w_1w_2 \cdots w_{k-1}\}$ 和 $\{w_{k+1}w_{k+2} \cdots w_n\}$ 分别。请注意，考虑到在线评论的短语境风格，我们将评论作为一个完整的目标，而不是像参考文献[6]中那样将其分解成句子。向量的维数取决于复习的长度，即 n 。表示记录单词左上下文的向量 w_k 为 $v_{left}(w_k) = (v_{leftk1}, v_{leftk2}, \dots, v_{leftkN})$ ，记录右侧上下文的向量为 $v_{right}(w_k) = (v_{rightk1}, v_{rightk2}, \dots, v_{rightkN})$ ，其中下标 i 表示 i th 向量的分量。 v_{wki} 表示 w 之间的共现强度 k 和 w_i 。为了计算我们将很快讨论验证的可能性距离，所有的向量必须以概率表示的形式。 $N \leq n$ 是上下文向量的维数，它对复习 R 的上下文中不同单词的数量进行计数。

重量 v_{wki} 是由：

$$v_{wki} = \sum_j e^{-\frac{|j-k|^2}{2\sigma}}$$

它同时考虑了空间因素和距离因素。这里我们使用一个高斯函数来测量距离，因为我们假设距离越近，它获得的权重越大。然后将其归一化，使其成为概率表示：

$$v_{wki} \leftarrow \frac{v_{wki}}{\sum_i v_{wki}}$$

然后将候选词或种子词转化为概率向量表示。然后我们将使用 Kullback-Leibler (KL) 距离^[8](记为 $KL(\cdot, \cdot)$)来度量向量 $v(c)$ 之间的距离 $i = (v(c_i)_{left}, v(c_i)_{right})$ 的 th 候选词和

$\mathbf{v}(seed_j) = (\mathbf{v}(seed_j)^{left}, \mathbf{v}(seed_j)^{right})$ of the j^{th} seed word, that is:

$$\mathbb{D}(c_i || seed_j) := KL(\mathbf{v}(c_i) || \mathbf{v}(seed_j)) = \sum_{k=1}^N P(d_k | c_i) \log \frac{P(d_k | c_i)}{P(d_k | seed_j)}$$

在 $P(d_k | i)$ 和 $P(d_k | \text{种子 } j)$ 表示 k 的概率权重 th c 的左(或右)上下文向量的维数 i 和种子 j ， N 为上下文概率向量的维数。由于 KL -距离的非对称性，即 $\mathbb{D}(c_i || seed_j) \neq \mathbb{D}(seed_j || c_i)$ ，我们采用如下对称距离(记为 $SD(\cdot, \cdot)$):

$$SD(seed_j, c_i) = SD(c_i, seed_j) := \mathbb{D}(c_i || seed_j) + \mathbb{D}(seed_j || c_i)$$

$\mathbb{D}(c_i || seed_j)$ 和 $\mathbb{D}(seed_j || c_i)$ 不仅使得 $SD(\cdot, \cdot)$ 具有对称性，而且还决定了它们的左右上下文分布。然后我们将定义候选词 c 之间的相似性(表示为 $SI(\cdot, \cdot)$) 种子字种子 j 基于对称距离，由：

$$SI(c_i, seed_j) := \frac{1}{1 + SD(c_i, seed_j)}$$

$SI(c_i, \text{种子 } j) \in (0, 1)$ 度量两个单词之间的相似度，两个单词的相似度越大。特别是， $SI(c_i, \text{种子 } j) = 1$ 当且仅当 $c_i = \text{种子 } j$ 。计算候选词与种子词的相似度后，我们将提出候选词 c 之间的相似度 i 五组中的一组是 G_j 由：

数学建模系列课程资料代码等请关注公众号：科研交流

$$SI(c_i, G_j) := \frac{1}{|G_j|} \sum_{seed_k \in G_j} SI(c_i, seed_k)$$

$SI(c_i | G_j) \in (0, 1)$ 度量候选词 c 的相似性 i 对 j th 一群感伤的种子话。综述 R 与 G 的相似性 j 被定义为：

$$SI(R, G_j) := \frac{1}{N} \sum_{i=1}^N SI(c_i, G_j)$$



$SI(R, G_j \in (0,1))$ 度量评审 R 与 j 的相似性 th 一组感伤种子词，越大越相似。CE 块的输出是一个 5 维概率向量 $CE(R) = (CE)R_1, \dots, CER_5)T \in R_5$ 。jth 组件 CER_j 强度与 G 组有关吗 j ，定义为：

$$CE_R^j = \frac{SI(R, G_j)}{\sum_{i=1}^5 SI(R, G_i)}$$

3.2.3 维德块

VADER^[7]是一个简单的基于规则的一般情绪分析模型，特别是社交媒体文本风格。基于人工生成的标准词汇特征^[7]列表，VADER 不需要任何训练数据，这显示了它扩展到情绪分析广泛任务的潜力。因此，在本文献中，我们基于我们生成的种子词扩展了良好标准的词汇特征列表，以处理评论上下文的情绪分析。此外，我们还将原来的四组预测结果扩展到五组预测结果，即，强阳性，弱阳性，中，弱阴性，强阴性在我们的 CE 区块一致性中。有关 VADER 框架的更多细节，请参考 ref^[7]。扩展的金标准列表附在附录 a。维德块的结果也是一个带有 j 的 5 维概率向量 th 分量是对应于 j 的强度 th 种子字群 G_j 。表示复习 R 的结果向量为 $VADER(R) = (VADER)R_1, \dots, VADERR_5)T \in R_5$ 。

3.2.3 提出 CE-VADER 进行情绪分析

我们提出的 CE-VADER 混合模型由两个 block 组成:CE block 和 VADER block。两个区块都将提供一个 5 维概率向量 $CE(R)$ 和 $VADER(R)$ 来回顾上下文 R ，每个组件是其对应组 G 的强度 j 。我们将使用这两个向量的平滑凸线性组合作为 CE-VADER 模型的最终强度概率向量(记为 $INT(\cdot)$)，即：

$$INT(R) := softmax(\lambda CE(R) + (1 - \lambda)VADER(R))$$

其中， λ 为融合系数，在本文献中，我们设置了 $\lambda = 0.5$ 来相等地衡量两个块的结果。我们使用 $softmax(\cdot)$ 作为平滑函数，其定义如下：

$$softmax(x_1, x_2, \dots, x_s) = \frac{1}{\sum_{i=1}^s e^{x_i}} (e^{x_1}, e^{x_2}, \dots, e^{x_s})$$

我们的经验结果表明，经过平滑后，强度向量将显示出很强的一致性星的评级。

本文提出的 CE-VADER 在 review R 上的情绪分类结果为 G 的组名 j 与组件关联的 $INT(R) = (INT)1R, \dots, int5R)$ 的最大值，由 $j_0 = \operatorname{argmax}_j \{int_j R\}$ ，其对应强度为 $int_{j_0} R$ 。

3.3 基于文本和基于评级的措施相结合

数学建模系列课程资料代码等请关注公众号：科研交流

唯一的评审 id 被表示为 id 。评审标题为 $Rhid$ ，审查机构 $Rbid$ ，星级 sid ，审查日期 id ，有益的投票 $hvid$ ，产品名称 $ptid$ ，产品编号 $pid \in \{B, M, H\}$ (B, M, H 代表婴儿奶嘴,微波,分别和吹风机)都与下标审查相关 id 。由于强烈的相关关系(总差异发生租赁超过 0.01%)之间产品称号,产品父母和产品 id ,即一旦其中一个是鉴于我们几乎可以唯一地告诉另外两个,我们只使用产品标题描述了产品在这个文学。我们不考虑市场，因为他们都来自美国。表示 $P(id)$ 对 = $(R)id$,年代 id ,高压 id , $idptidpid$), $Rid = (Rhid, Rbid)$ 是整个复习文本，包括标题和正文。



我们提出了重要性率(表示为 IMP)的每一个审查, 以其有益的投票, 对应的星级措施 $VEC(s_{id})$ 和基于文本的度量 $INT(R_{id})$ 检查文本的清晰度。它由以下公式定义

$$IMP(id) := (1 + hv_{id}) \cdot \exp[-\alpha(1 - \frac{INT(R_{id}) \cdot VEC(s_{id})}{\|INT(R_{id})\| \|VEC(s_{id})\|})] \cdot \exp[\beta(\sum_{i=1}^5 int_R^i \log(int_R^i))]$$

基于文本的测量和评级之间的余弦距离: $1 - \frac{INT(R_{id}) \cdot VEC(s_{id})}{\|INT(R_{id})\| \|VEC(s_{id})\|}$

基于测度, 计算两者之间对应的保真度, 零是其关联权值。 $-\sum_{i=1}^5 int_R^i \log(int_R^i)$

是文本上下文的熵, 越低越清晰, $\beta > 0$ $i=1$ 为关联权值。重要性越高, 评论文本和评分的信息量就越大

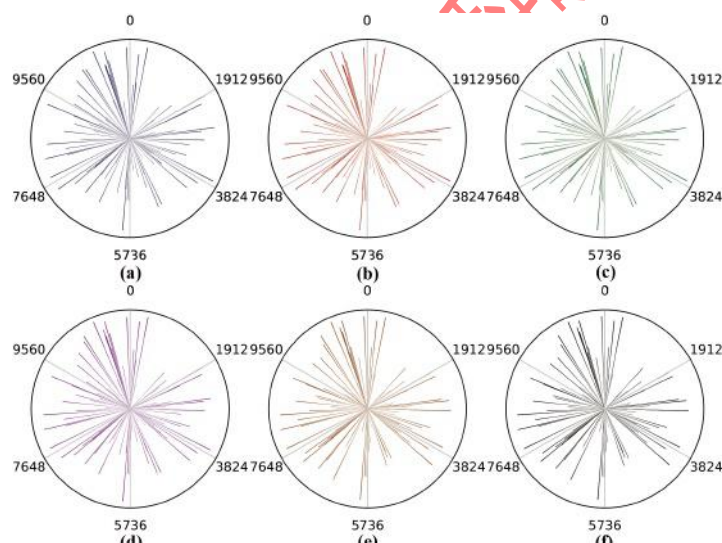


图 4:在“吹风机”数据集上实现了我们所提议的模型。

我们演示了由不同参数(如, 线性线越长, 我们的模型显示的信息就越多。(人事部, 人事部)的设置为:a (1,1);b (1, 3);c(1, 5);d(2, 1);e(4,1);f(5, 5)。如图所示, 我们的模型显示了这两个参数的鲁棒性。

3.4 模型实施、敏感性分析及结果

我们针对给定的数据实现我们提出的模型。我们设置了参数 α 式(1)中 $\alpha=1$ 。3.3 节提出的模型包含两个参数:首先分析这两个参数的灵敏度。

如图 4 所示, 我们在“吹风机”数据集上实现了我们的模型, 由我们的模型显示的最具信息性的前 1%的综述显示了对两个参数的超强鲁棒性。我们还使用 DTW 相似性来量化我们的模型在两个参数上的鲁棒性, 如图 5 所示。DTW 相似度越小, 两个排名越相似。读者可以参考 ref.^[9] 获得更多关于 DTW 相似性的细节。我们设 $cyp = 1$, $cyp = 1$ 为基线。计算 DTW 相似度, 最大值为 14.3, 在 11470 条评论的情况下, 这个值比较小。再一次, 我们的模型显示出了较强的健壮性。

附录 C 中附有三种产品的排名所列出的前 1%最具信息的评论和相关评级。

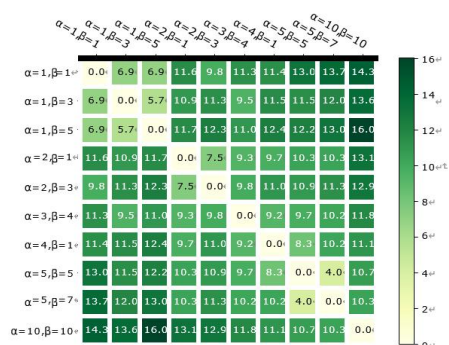


图 5:我们的模型在“电吹风”数据集上实现的 DTW 相似性

我们的模型在“电吹风”数据集上实现的 DTW 相似性，在不同的数据集上，and 的值不同。我们设置了以下参数:1 = 1,2 = 1。最大相似度为 14.3，考虑到我们总共有 11470 篇评论的情况，这个值很小。它显示了我们模型的健壮性。

4 用差分方程测量基于时间的图案

在本节中，我们建立了一个基于差分方程的模型来表示产品声誉的变化。本节的其余部分安排如下。在 4.1 节中，我们详细阐述了我们的模型。在 4.2 节中，我们分析了模型的敏感性，并在三种产品上实现了模型。

4.1 基于差分方程的模型

我们建议用“声誉”来表示产品 P 在 t 时期的声誉，用 $\text{REP}_{t_0, \theta}(T, P)$ 表示。但是，我们假设产品的声誉会逐渐改变。由买方的审查文本和星级费率。我们将使用一个差分方程来表示声誉。声誉的差异(例如:声誉增长率)可以表示为:

$$\Delta \text{REP}_{t_0, \theta}(T, P) := \frac{1}{2z} \left(\sum_{T-t_0 \leq r_{id} \leq T, p_{id}=P} \text{IMP}(id) \cdot (\theta s_{id} + (1-\theta) \text{argmax}_j \{ \text{int}_{R_{id}}^j \} - 3) \right)$$

其中 Z 为归一化常数，定义为:

$$Z = \sum_{T-t_0 \leq r_{id} \leq T, p_{id}=P} \text{IMP}(id)$$

为简单起见，我们对 ΔREP 的值进行规格化 $t_0, \theta(T, P)$ 到 $[-1, 1]$ ，与负面感觉或一星、二星评价相关的负值，与正面感觉或四星、五星评价相关的正值。

参数 $\theta \in [0, 1]$ 是 3.2 节中我们提出的 CE-VADER 模型对星星评级和基于文本的度量的权重系数。我们假设产品的运输需要一些时间，而一些客户更喜欢在购买后的一段时间内进行评审。因此，我们认为所有的 id 都满足 $T-t_0 \leq r_{id}$ 时间标签 $\leq T$ ，其中 T_0 是一个门槛。请注意，买家有将近 90 天的时间来留下反馈，因此 $t_0 \leq 90$ 。在本文献中，我们设置了等权重的星级率和综述上下文的 $\text{lai} = 0.5$ 。我们设置 $t_0 = 10$ 是这篇文献。那么声誉可以用以下差分方程表示:

$$\text{REP}_{t_0, \theta}(T, P) - \text{REP}_{t_0, \theta}(T-1, P) = \Delta \text{REP}_{t_0, \theta}(T, P) - \text{PEN}(T, P)$$

钢笔(T, P)是惩罚因素，用来计算低星级评级或负面评论对声誉的破坏程度，因为我们假设买家会更关注那些负面评论。它的表述如下。

$$\text{PEN}(T, P) = k_2 \times \text{sigmoid}(k_1 \times \text{REP}_{t_0, \theta}(T-1, P)) \times \frac{\#\{id | \theta s_{id} + (1-\theta) \text{argmax}_j \{ \text{int}_{R_{id}}^j \} \leq k_3\}}{\#\{id | T-t_0 \leq r_{id} \leq T, p_{id}=P\}} \quad (16)$$

k_1, k_2, k_3 为惩罚因子的阈值。我们集 $k_3 = 2$ 作为量化的阈值“负面评论”。因此，“负面”的传播

$$\frac{\#\{id | \theta s_{id} + (1-\theta) \text{argmax}_j \{ \text{int}_{R_{id}}^j \} \leq k_3\}}{\#\{id | T-t_0 \leq r_{id} \leq T, p_{id}=P\}}$$



评论定义为 $\text{sigmoid}(x) = \frac{1}{1+e^{-x}} \in [0,1]$ 。我们设置 sigmoid 项目是为了亲密的社会行为，一旦产品获得了良好的声誉，人们会更加重视“负面评价”。 k_2 衡量惩罚因素与增长率之间的比较。

4.2 模型实施，敏感性分析和结果

在 4.1 中我们构建的模型中，有两个可调参数，即 k_1 和 k_2 。我们在三类产品的数据上实现我们提出的模型。并显示我们的模型对这两个参数的灵敏度。图 6 描述了参数设置为 $k_1 = 0.5$ 和 $k_2 = 20$ 。如图所示，“微波”的名誉值为负值；而“安抚奶嘴”赢得了良好的声誉。

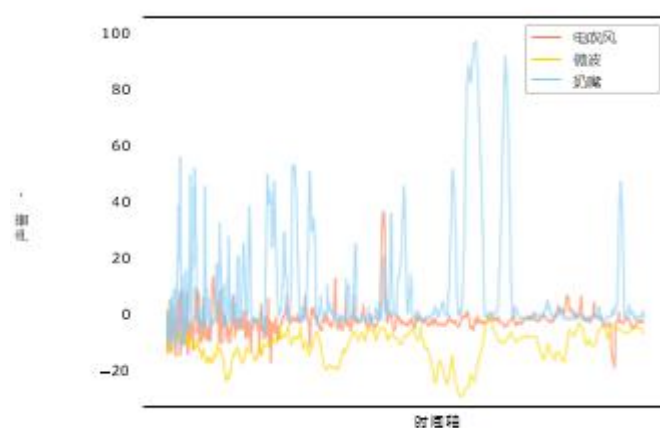


图 6:带 k 的产品声誉曲线 $k_1 = 0.5$ 和 $k_2 = 20$ 。

名誉率下降。然后，分析了模型对两个参数的敏感性。 k_1 和 k_2 。我们的模型在 k 上显示出极强的鲁棒性 1 以及对 k 的敏感度 2 在图 7 中。如图 7-B 所示， k 的不同设置 2 将导致“安抚奶嘴”数据上的口碑率增长率不同，但趋势相同。但是， k 的小值 2 将导致一个指数大的声誉增长率。正如我们之前讨论过的， k 的小集合 2 表示较少注意负面评论。

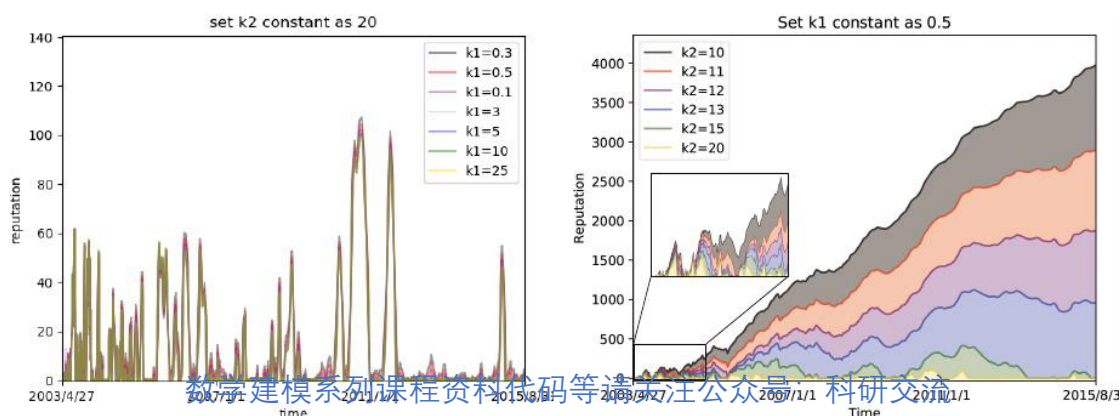


图 7:参数 k 的灵敏度分析 1 和 k_2 。

A 我们的模型对 k 表现出极强的鲁棒性 1。B 我们的模型显示了对 k 的敏感性 2。

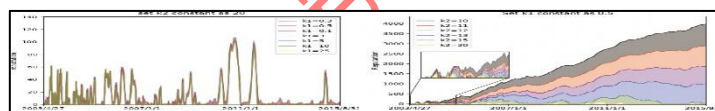


5 预测潜在的成功或失败

在最后一节中，我们构造了一个映射 $REPt00: [T0T1[B, M, H] \rightarrow R^+$ 在 $[T0T1]$ 是给定审查日期的时间范围。在本节中，我们将使用时间序列预测方法来预测 REP 的值 $t00$ 在扩展的未来域上，即 $[T1T2[B, M, H]$ 为某 $T2 > T1$ 在 5.1 节。根据预测的 $REPt00$ 在未来的时域中，我们将在 5.2 中评估每个产品的成功或失败潜力。在第 5.3 节中，我们在真实数据上显示结果并预测其潜在的成功或失败。

5.1 预测未来声誉的时间序列

本文采用自回归模型(AR)作为时间序列预测方法，预测未来的商誉率增长或下降。AR 模型可以用以下差分方程表示：



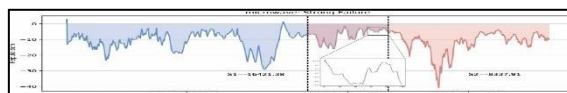
在 ε_T 就是白噪音。AR 模型去搜索系数 $(a_0, \text{一个 } 1, \dots, a_p)$ ，在时域拟合给定数据 $[T_1, T_2]$ 的最小均方根误差(RMSE)。

5.2 评估成功或失败的可能性

我们从 T 定义产品 P 的平均信誉 $0 T1$ 为：

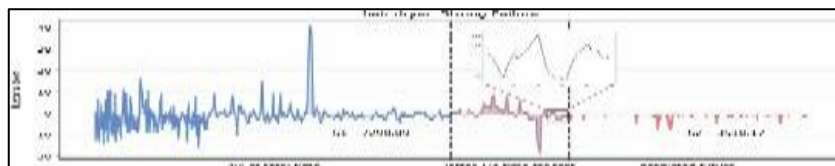


同样，我们可以定义产品 P 在未来时域的预测平均信誉 $[T]1, t]$ 表示 $t < t_2$ 基于 5.1 节



构建的模型，如下所示。

将给定数据计算出的平均声誉与 Eq.(18)和 Eq.(18)预测出的平均声誉之比定义为声誉变



化率。定义为：

我们提出一个模糊评估系统来预测产品 p 的成功或失败。系统的总体架构如图 8 所示。我们将结果分为四类，即“强成功”、“弱成功”、“弱失败”、“强失败”。

5.3 模型实施与结果

在 Eq.(17)中，我们对 $p = 50$ 的给定数据执行我们的模型。我们设置 34 个数据在给定



的时域([T0T1])。[T0, 43 T1 + 14 T0]拟合回归系数(a0,一个 1, ..., ap)。

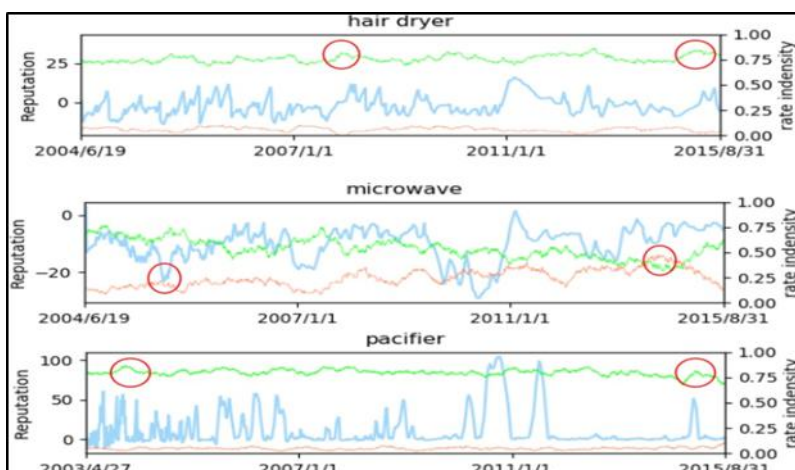


图 8:我们提出的用于预测每个产品潜在成功或失败的模糊评估系统的总体架构。

后 14 个评估 AR 模型，即。在时域[34 T1 + 14 T0T1]。在评价领域，我们发现该模型能很好地拟合和预测“信誉”率，最大 RMSE 率为 0.031。我们评估域[T]上的潜在成功或失败 1T2)和 $T2 - T1 = 21 (T1 - T0)$ 。奶嘴、微波炉、吹风机的预测口碑率分别如图 9、10、11 所示。根据第 5.2 节中提出的框架，它们被预测为“弱成功”、“强失败”和“强失败”。

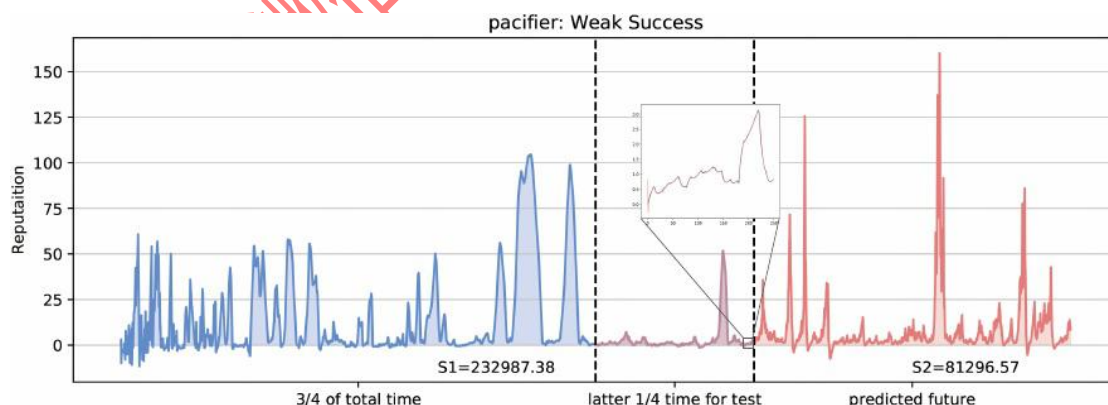


图 9:“安抚奶嘴”数据的模型实现。预计会有微弱的成功。

6 具体评级和描述符分析

在本节中，我们在第 6.1 节和 6.2 节中对具体的评分和描述符进行了重复的分析。

数学建模系列课程资料代码等请关注公众号：科研交流



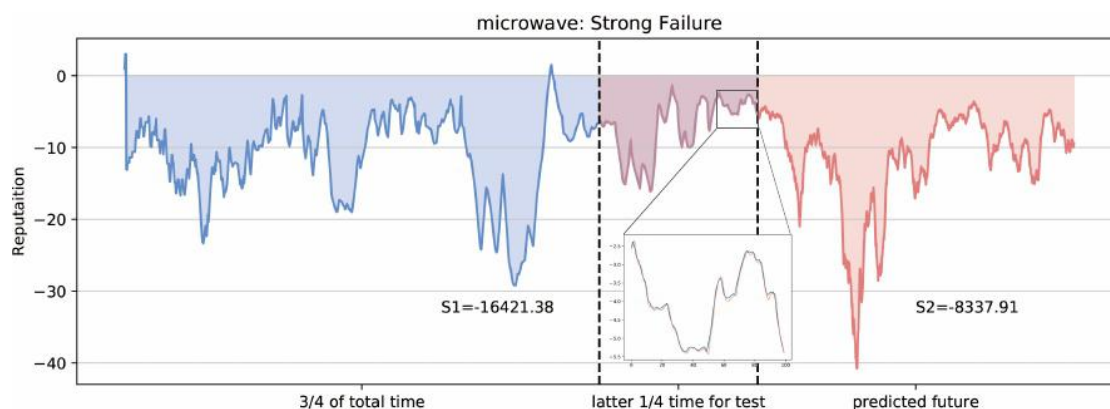


图 10:“微波”数据上的模型实现。预计会是一个巨大的失败。

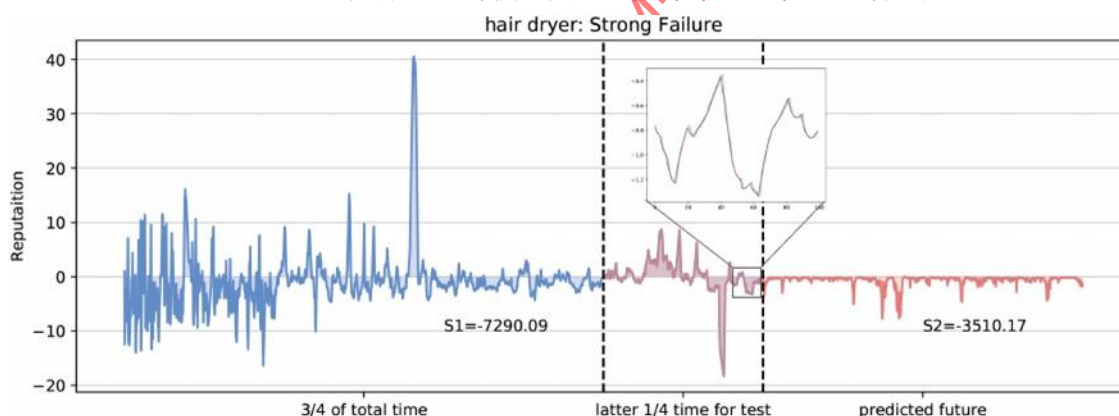


图 11:“吹风机”数据的模型实现。预计会是一个巨大的失败。

6.1 具体星级评级与评级频率的相关性

在这一节中，我们首先分析了特定的星级评级与商誉率，即商誉率之间的关系。一星评价的负面情况较强，五星评价的正面情况较强。我们采用我们提出的信誉率作为基于文本的度量。如图 12 所示，我们可以观察到五星评价比例的增加导致了“声誉”率的瞬间提升。而在五星级评级比例达到顶峰后，口碑率立即下降。声誉率的增长也与一星评级比例的下降有很强的相关性。

然后，我们评估星级评级与图 13 中有效投票之间的相关性。在所有类型的数据中，与其他四个评级相比，五星评级再次显示出与有效投票的强烈相关性。婴儿奶嘴，微波炉和吹风机。

此外,以前的工作[10]的研究也指出,人们认为极端情况下在所有网上评论(正面和负面)更多的有用的信息,而不是温和的例如,审查文本分类为“强阳性”或“强烈的负面”和一颗或五星级率问题。

因此，在本文献中，我们将重点放在极端评级的影响上。



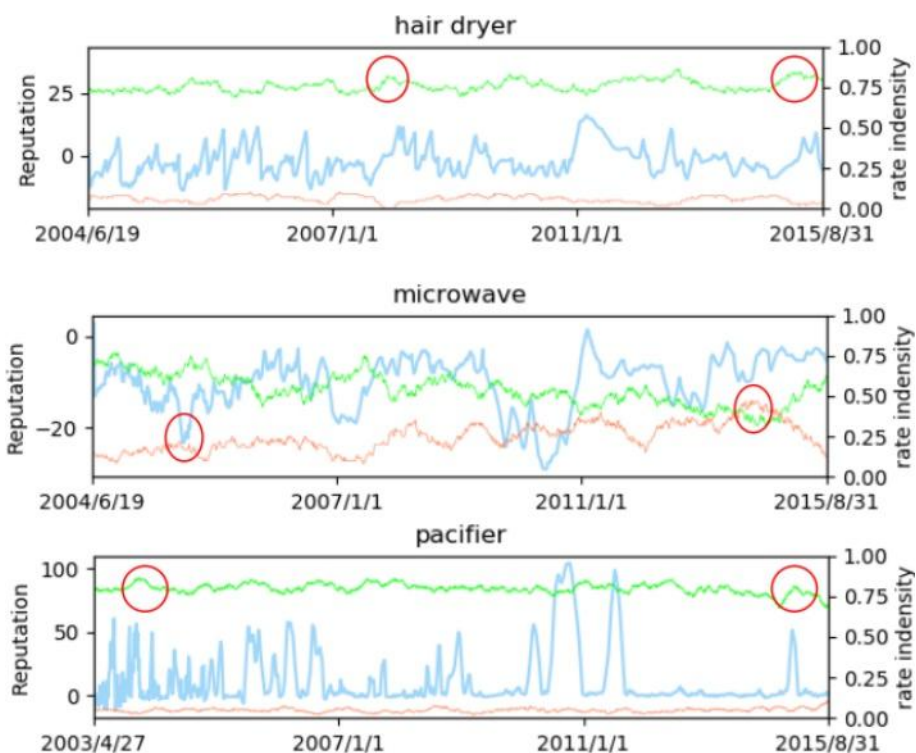


图 12:“声誉”率与一星或五星评价比例的相关性。

蓝线表示信誉时间序列，绿线表示五星评价比例，橙色线表示一星评价比例。红色的圆圈突出了典型的例子，即较高的五星率会立即导致声誉率的提高，而一星率则会下降。即是一星和五星的评级，而不是 2 星到 4 星的评级，基于上述三个事实。

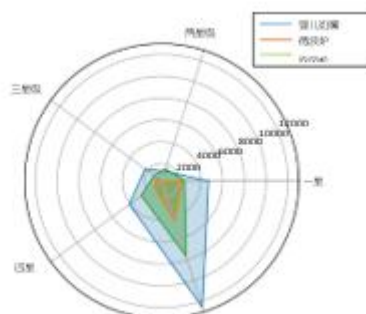


图 13:星级评级与有用投票之间的相关性。

通过验证 1 星或 5 星产品相关的评论文字词频的模式，我们发现消费者不太可能购买那些与低-相关的产品他们更愿意购买那些高星级的产品。我们采用一个特定的案例，即连续出现与一星或五星利率相关的 x 评论。以受试者为基线，吹风机 $x = 10$ ，奶嘴 $x = 5$ ，微波炉 $x = 5$ 。由于与其他两种情况相比，微波的低五星率和坏名声(正如我们之前在 4.2 节中讨论的)，我们专门为它的低五星率设置 $x = 5$ 。实验结果表明， x 越小，频率越高， x 越大，频率越低。为了简单起见，我们如上所述设置了一个中等的 x 。

如表 2 所示,更多的评论被煽动五星级评级连续出现后,在这两种情况下,拥有一个好的或坏名声(声誉是由我们的模型制定在第四节)。这表明,人们更愿意看到后产品连续发生大星级评级。而在连续出现一星评级后，人们对于声誉不好的产品，即产品的评级积极性降低了。微波炉和吹风机。看到连续出现一星评级，人们就会被强烈的激励去给信誉好的产品打分。



产品	额定频率 在一系列的 五星级评级(天)	额定频率 在一系列的 一星评级(天)	整体 评级频率(天)	声誉
婴儿奶嘴	1.9444	1.56000	2.37828	好
微波炉	8.66667	30.61904	24.39975	坏
吹风机	1.58207	16.37500	4.29293	坏

表 2:评分频率与一系列极端评分案例的比较

总体评分频率与婴儿奶嘴、微波炉、吹风机的平均声誉数据。频率越小，购买者对产品的评价就越频繁。五星评价的连续出现总是激励人们去打分。连续出现的不良产品一星评级，导致频率大大降低，同时强烈的鼓动人们投票给良好的信誉产品。

数字表示 14 为 2012 年 8 月 31 日至 2015 年 8 月 31 日，吹风机评分总量发生显著变化时，每个月的星级评分总数以及一星和五星评分数。我们可以观察到五星评分的集中总是伴随着评分总数的增加，而一系列的一星评分则会导致评分的下降。

6.2 特定质量描述符与等级的相关性

6.2.1 提出 CE-VADER 进行情绪分析

有些词通常与导致不同评分等级的情感感受紧密相关，例如，“热情”带有积极和高兴的感觉可能会导致较高的评分，而“失望”带有消极的感觉。在本节中，我们使用朴素贝叶斯模型来确定特定单词是如何与评级等级相关联的。具体单词 w 的 star-rating s 的朴素贝叶斯分布概率如下：

$$P(s|w) = \frac{P(w|s)P(s)}{P(w)}$$

通过对概率 $P(s|w)$ 的评估，我们可以衡量特定单词 w 与星号率 s 之间的相关性。

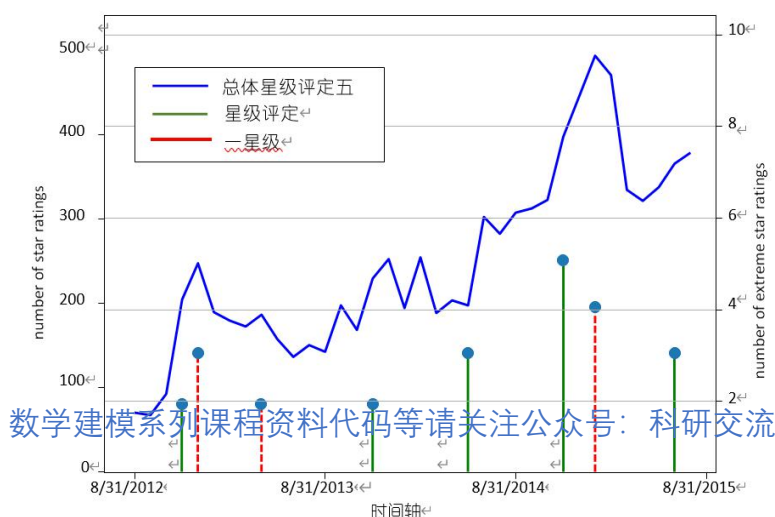


图 14

从 2012 年 8 月 31 日到 2015 年 8 月 31 日，电吹风数据上的一星或五星评级与整体星级评级之间的相关性的典型例子。



6.2.2 模型实施与结果

$P(s|w)$ 越高，单词 s 与星级 s 的相关性越强。在提供海量数据的情况下，我们使用频率来近似 Eq.(21)中的概率，即 $P(s) = \# \{id|sid=s\}$ 。我们展示了与五星评分和一星评分相关的十大单词。

表 3:与五星评分和一星评分相关的前十个单词。

与五星级评级密切 相关的词	$P(s = 5 w)$	与 1 星评价强烈相 关的词	$P(s = 1 w)$
容易	0.8518	或	0.1416
爆米花	0.8333	出	0.1434
爱	0.7809	不	0.1483
功能	0.7778	没	0.1513
按钮	0.7692	背部	0.1546
大的	0.7685	热	0.1718
多	0.7452	因为	0.1759
高	0.7250	月	0.1940
以	0.7200	关	0.2083
也	0.7196	烘干机	0.2125

7 设计特点的吸引力分析

在本节中，我们从前 1%最具信息性的评论中提取描述婴儿奶嘴、微波炉和吹风机性能、外观等的关键词。通过计算这些关键词在所有评论中出现的频率，并分析信息最丰富的 1% 评论的内容，我们将一些对消费者有吸引力的设计特性，以及关键词的组和关键词出现的数量显示在附录 B 中。

1.婴儿奶嘴:消费者最关心的是婴儿奶嘴的尺寸、外观、方便性和安全性，见表 4。人们更关心它对婴儿是否合适和安全，所以各种尺寸和安全材料的选择是必要的。此外，可爱的图案很受婴儿的欢迎。

2.从表 5 可以看出，消费者最关心的是微波炉的外观、价格和成分。从表 5 可以看出，人们最关心的是微波炉的性价比。许多评论欢迎多功能但昂贵的微波炉。在产品价格昂贵的情况下，人们更愿意拥有尽可能多的功能，比如旋转烤架、精确的计时器和多种使用模式。

3.吹风机:从表 6 可以看出，吹风机的功率和外观是消费者最关心的，说明消费者更喜欢功率大的便携式吹风机。同时，安全性和工作容量影响用户体验，如何在增加的功率和安全安静的运行环境之间保持平衡，对阳光公司来说意义重大。特别是一些特别的设计可能会刺激消费者购买，例如可折叠的把手可改善吹风机的携带性。

8 销售策略和建议

在本节中，我们基于前几节对婴儿奶嘴、微波炉和吹风机这三种产品的结果，为阳光公司提出了自信的在线营销策略和建议。我们最自信的建议包括具体的理由(给阳光公司)如下：

- (一般性建议)有高帮助投票和与强烈感情相关的评论通常是信息丰富的。实时收集这些评论和星级评价。注意基于文本的评论和星级评分一样重要。

- (一般推荐)始终保持较高比例的五星评价，因为它与产品的声誉有很强关系。



•(一般建议)如果产品口碑不好,比如低星级的评价比例高,则更注重负面评价或低星级评价。并立即做出反馈和调整,因为这些评论或评级会继续破坏产品的声誉。大多数人谁评级一星状态,他们将更新我的评论,一旦他们听到或不 head 从 com-公司。大约有 30%的一星评级与保修服务不及时或来自公司的反馈有关。

•(吹风机的具体理由)吹风机的条件似乎很差。在我们评估的前 100 个最具信息量的评论中,有 13 个评论报告了吸烟、扔火花或坏掉的情况,这些可能会导致一些危险。由于吹风机的口碑下降,我们建议您在口碑上升后再投放市场,这样有利于产品的顺利启动。

•(微波的具体理由)微波在三种产品中声誉最差。它拥有最高比例的一星评级和强烈的负面评论。在信息量最大的 100 条评论中,人们普遍抱怨恶劣的环境和潜在的着火危险。我们的模型表明它可能会失效。因此,我们建议等到微波的声誉恢复和保持稳定。

•(奶嘴具体理由)奶嘴口碑好,正面评价和五星评价比例最高。根据我们的模式,它也将获得成功,所以现在是把它推向市场的好时机。然而,在信息最丰富的 1%评论中,他们报告了漏水的情况。企业应注意连接或联合部分。

•(一般建议)奶嘴、微波炉、吹风机等产品成本低,同时对消费者的专业要求也较低。产品信息越完整,买卖双方信息不平等造成的损失就越小。当然,信息的真实性需要得到保证。

•(一般建议)我们的研究表明,一系列的五星评级可以激发消费者的评论。因此,当您的产品有更多的五星评价时,我们建议您加大促销力度,形成积极的反馈。

•(一般建议)创意和吸引人的广告将增加产品销售。在线市场让消费者更容易浏览商品的多种替代品,这导致了公司之间的激烈竞争。设计有趣、醒目的广告可以有效提高产品的竞争力,所以阳光公司应该在广告上进行适当的投资,以获得更多的利润。信息量最大的前 1%的评论表明,一些消费者非常关注该公司的网站。

9 优点和缺点

9.1 优势

1.新鲜事物。据我们所知,我们是第一个提出 CE-VADER 混合模型的人,该模型用于在在线市场上审查基于文本的情绪评估。

2.准确性。在验证时间上,50 订单 AR 模型的最大 RMSE 为 0.031 域。基于文本的度量与基于评级的度量很好地关联在一起。

3.泛化。我们提出的框架可以自由地应用于任何数据集,例如,来自任何在线平台的任何产品的评论和星级评级。

4.鲁棒性。我们的模型对大多数参数显示了很强的鲁棒性。

9.2 缺点

数学建模系列课程资料代码等请关注公众号：科研交流

1.耗时的手工注释。手工注释从 CE-VADER 模型的评论生成的种子非常耗时。

2.CE-VADER 模型无法识别同一词的不同形式,具有特殊的变异规律。CE-VADER 模式不能识别动词的过去式、名词的复数和具有特殊变化的形容词的比较级形式,如 is (vs . was)、children (vs . child)和 better (vs . good)。



3.遗漏了其他潜在的相关因素。在分析具体的评级和描述符时，我们并没有考虑到类似亚马逊的促销营销策略。

10 结论

为了破解亚马逊评分和评论的秘密，我们提出了一系列新颖的模型来解决从选择信息最丰富的评论到识别评论质量描述符的子问题。该模型具有较高的精度和鲁棒性。

1.信息评估模型可以将基于文本的度量与基于评级的度量相结合，其中我们提出了一种新的 CE-VADER 混合模型作为基于文本的情感分析度量。我们可以根据所提议的模型对每个评论和评级的信息量进行排序。信息率与有用的投票相关。更具体地说，帮助投票越多，他们拥有的五星评级就越多，评论机构的时间越长，他们就越有可能被评价为有信息的评论。但是，请记住，评价不高、信息熵大的评论，既有正面评论，也有负面评论，加上“然而”、“但是”等字样，同样具有很大的参考价值。

2.利用差分方程模型构建“信誉率”来量化婴儿奶嘴、微波炉和吹风机三种产品的信誉。婴儿奶嘴有正面的评价，吹风机有负面的评价，微波炉的评价最差。通过改进的 AR 算法，对这三种产品的未来信誉趋势进行了预测。

3.在分析星级评价和特定词的分布时，我们使用连续的极端评价和一组特殊词来识别特殊的评论描述符。连续的极端评分对总销量有明显的影响，特殊词语的出现可以判断较高概率的评论评分。

给阳光公司市场总监的一封信

尊敬的阳光公司营销总监：

根据贵公司的要求，我们分析了贵公司拟引进销售的婴儿奶嘴、微波炉、吹风机等竞品在亚马逊上的评分和评论。我们形成了三种模型，分别为信息评价模型来评估每次评审的信息量，基于差分方程的模型来制定产品声誉的变化，基于时间序列预测的模型来预测产品的潜在成功或失败。我们得到了一些有意义的结果，这有助于吸引消费者和制定适应性的网上销售策略。

首先，我们构建了信息性评价模型，帮助您方便地跟踪信息性评论和评分。该模型可以根据星级评级、评论文本和有用的投票来评估每个评论是否具有信息性，然后根据信息量对评论进行排序。我们相信，当您处理大量的评论信息时，这种信息性的评价对您的工作至关重要，因为信息性的评价通常为设计特性提供更有建设性的输入，因此在声誉分析中更有参考价值。我们通过筛选婴儿奶嘴、微波炉和吹风机的前 1% 最具信息的评分和评论来了解消费者的喜好和关注。经我们分析，目前比较流行的设计特点如下：

- 婴儿奶嘴：需要多种尺寸和安全材料的替代品。此外，可爱婴儿喜欢图案。
- 微波炉：多功能微波炉，配有旋转烤架、精确定时器、多种使用模式等，深受消费者欢迎。
- 吹风机：消费者更喜欢小功率的吹风机（在安全的前提下）。一些额外的设计，如可折叠的手柄，有利于吸引消费者。

然后我们建立了基于时间的措施来预测在线市场中产品声誉的增加或减少。我们的模型可以准确预测产品在未来很长一段时间内的声誉变化。这种精确的预测有助于你在声誉下降之前制定策略。



数学建模系列课程、资料、代码等请关注微信公众号：科研交流

在声誉预测的基础上，我们规定未来的声誉增加或减少的程度，作为对产品潜在成功或失败的判断。我们的分析显示，总的来说，婴儿奶嘴是潜在的成功产品，而微波炉和吹风机失败的风险更大。

根据我们的分析结果，我们为贵公司制定了合理的销售策略：

- (1)我们建议贵公司在声誉上升时，将微波炉和吹风机投放市场。当婴儿奶嘴的名声在上升的时候，最好现在就把它们放到市场上。
- (2)产品信息越完整，买卖双方信息不平等造成的损失就越小。
- (3)当您的产品有更多的五星评价时，我们建议您加大促销力度，形成积极的反馈。
- (4)当你的产品声誉下降时，关注一星评级和评论。

感谢您在百忙之中阅读我的来信。

数学建模系列课程资料代码等请关注公众号：科研交流



参考文献

- [1] Peter Evans 和 Annabelle Gawer。平台企业的崛起:全球调查。2016.
- [2] Seyed Pouyan Eslami, Maryam Ghasemaghaei 和 Khaled Hassanein。哪些在线评论对消费者最有帮助?方法调查。《决策支持系统》, 2018 年, 113:32 - 42。
- [3]理解在线评论帮助性的决定因素:一项元分析调查。决策支持系统, 102:11,2017。
- [4]胡亚涵和陈宽津。预测酒店评论的实用性:评论可见度的影响, 以及酒店星级和评论评级之间的互动。国际信息管理杂志, 36(6,Part A):929 - 944, 2016。
- [5] Y 陈和 J.Xie。在线消费者评论:口碑作为营销传播组合的新元素。管理科学, 54(3):477-491,2008。
- [6]余良智, 吴强龙, 张培灿, 朱宣寿。利用上下文熵模型对股票市场新闻情绪分类中的情绪词及其强度进行扩展。基于知识的系统, 41:89 - 97,2013。
- [7]克莱顿 J 哈托和埃里克吉尔伯特。Vader:一种简洁的基于规则的社交媒体文本情绪分析模型。2014 年第八届国际 AAAI 博客与社交媒体大会。
- [8]所罗门 Kullback。信息理论和统计。约翰·莱利和儿子们。公司。纽约,1959 年。
- [9]我同意, Marc Wichterich, Ralph Krieger, Hardy Kremer 和 Thomas Seidl。预期 dtw 用于时间序列数据库中有效的相似度搜索。VLDB 捐赠程序, 2(1):826-837,2009。
- [10] Sangwon Park 和 Juan L. Nicolau。网上消费者评论的不对称效应。旅游研究年鉴, 50:67 - 83,2015。

附录:

附录 A 注释种子词和频率

我们手工将种子词注释为五组, 如下所示。

1.强烈正面评价:棒(3006)、推荐(1174)、完美(761)、最佳(656)、最爱(368)、完美(183)、棒极了(125)、优秀(114)。

2.弱阳性:像(3624),爱(2766),爱(2753),易(2154),(1424),(1286),(1054),可爱(1002),爱(876),(855),更好的(821),(738),(731),(687),喜欢(647)希望(644),(585),快乐(577),(521),更容易适合(395),(360),喜欢帮助(311),很高兴(229),清洁(193),可爱的(191),可爱(191)娱乐(181),爱(163),(157),(153),(152),帮助高兴(141),容易(131),(131),兴奋(118),舒适(114),温暖(114),爱(103),可爱!(99), enjoy (96), enjoy (86), cool(85)。

3.温和的:宝宝(4998),(3852),她(3678),(3537),(2913),(2658),(2446),(2239),儿子女儿(2114),(1867),(1184),她(763),(626),婴儿容易(619),礼物(556),(513),超(489)玩具(448),(437),(407),(400),不会小(302),(303),不会女孩》(293),(293),(282),(251),(218),把(213),重(210)、朋友(205),大(189),(171),品牌问题(169),大(167)人(166),(166)担心,附加(165),(160),快(159),更喜欢(158),失去了(157),保健(157),取代(157),(151),孙女(150),(146)问题,快速(142),便宜(142),女儿的(140)、安全(139)、小(136),(136),(131),(129),(128),喜欢(128 年),清洁(123),孔(123),礼物(120),很容易(118),简单(118),(113),便宜(110),(106),除非原(107),略(105),护理(103),礼物



(98),薄(94),侄子(94),保存(94),松(93),调整(92), 更好(92), 脏(87)。

4.弱负:不(1283),(1011),没有不能(780),(734),(354),(346)问题,不是(297),(285),不但是(252)困难(248),(244),不能(211),(175),(173),(149)问题,不是(143),(132),(126),(124),难度(118),废(118),错误(102),麻烦(101),疼痛(90),(180)。

5.强烈否定:但是(7128), 失望(170), 不是(90), 不是(113)。

附录 B 关键字在不同关键字组中出现的数目

表 4 在总共 18937 条评论中列出了婴儿安抚奶嘴的关键字出现次数。

表 4:婴儿奶嘴不同关键词组关键词出现次数

集团	关键字出现的次数
大小	3402
外观	3399
方便	3138
安全	2374
组件	706

表 5 在总共 1616 条评论中列出了 hair dryer 的关键词出现次数。

表 6 在总共 11470 个评论中列出了吹风机的关键字出现次数。

表 5:微波不同关键字组中关键词出现的次数

集团	关键字出现的次数
外观	2350
组件	2289
价格	2234
设置	1049
安全	617
权力	594

表 6:吹风机不同关键词组中关键词出现次数

集团	关键字出现的次数
权力	6774
外观	4656
特殊的设计	3973
头发质量	2666
安全	1544
工作容积	1542

附录 C 排名前 1%的最有信息的评级和评论

根据我们的模型预测的“吹风机”数据集的“重要性”率，我们列出了以下最具信息性的 1%评论及其相关评级。

1.<无问题>(5 星评级):优于预期。伟大的产品，卓越的交付。

2.>(5 星级):我是一个小肩膀问题，这是非常紧凑和轻的重量。它以强大的力量出击!

3.<好产品. .爱它!>(5 星):我的头发卷曲浓密，使用这个产品使它更容易打理。增添光泽和柔软…不是油性/重. .爱死它了。

4.<good purchase great quality>(5 星评级):我经常用它，我是一个发型师，用它来帮助我的卷发客户，它从来没有让我失望过。价格好，质量好，买的好。

5.<头发浓密，不用担心。>(5 星评级):喜欢这款烘干机，工作起来又快又好。给我妈妈买的礼物，她非常喜欢!她的头发很浓密，电吹风对她很有用。



6.<太好了!很适合我浓密的头发。>(五星级):我有一头浓密的齐肩长发,通常需要 35 - 40 分钟风干。用这台烘干机,我在 20 分钟内就完成了,我的头发看起来又顺又软。我对整件事持怀疑态度,但这对我的头发有很大的影响。可拆卸
过滤器是决定因素。我把干衣机扔了,就因为我洗不干净。优秀的产品!

7.<爱这干燥!>(5 星评级):作为一个专业的发型师,我已经厌倦了思的吹风机吹不干头发比便宜的任何。然后,我读了关于这台烘干机的评论,又一次我想,是啊。所以,我买了它,我必须说,这吹风机晃动。它可以让头发在一半的时间内变干,让头发光滑有光泽!最后,我找到了一个能兑现承诺的

8.<很棒的烘干机,很棒的价格>(5 星评级):很棒的烘干机,很好用。当我需要它的时候,它有很大的热量和体积,很好的冷却按钮。快吹干我的头发。我的头发在一夜之间就坏了,所以我只留了一天头发。还不错,因为我没有时间去实体店购物。

9.<卓越的价值!>(五星评级):这是如此的便宜,但产品不是-优良的质量!!喜欢它,经常使用它!

10.<优秀,强力行动>(5 星评级):我们已经连续使用这款吹风机几个星期了。总的来说,这是一个强大的吹风机,可以让你的头发快速干燥,并保持良好的热量平衡。没有便宜的烘干机可以开发的热点。是的,它的重量和体积都很大,但那是因为它的建筑质量。我毫不怀疑这台烘干机在未来几年还会继续使用。

11.<漂亮整洁的小烘干机>(五星级):我妻子在我们住的汽车旅馆用过一台。她很想要一个,所以我就买了。你的老妇人也会把它挖出来的,因为它很有力量,还有一个线轴。当你不得不拆开它,把她的头发从里面挖出来的时候,你会恨你自己,因为她抱怨它不像以前那样用力吹。但是女人都是那样的抱怨抱怨……如果你能在她需要它来擦干她的湿头之前把它拆开,清洗,再重新组装成工作状态,你就幸运了。我认为核弹更容易研制。

12.好产品,好价格。>(5 星评级):这是我的第二台相同型号的吹风机。(Conair Ionic Conditioning Pro Style 1875-Watt Hair Dryer)[[ASIN:B00005O0MZ]]我想再次使用它。

13.<爱它!!史上最好的烘干机>(5 星评级):它让我的头发很好,没有冻住!!我很伤心,我错掉了我的 3 年后,我有我的,但我将防御再买它!!

14.<好头发!>(五星级):我的头发很细,不喜欢别人给我做发型。使用这个烘干机,我有;

体积
机身
卷曲度

,我不需要使用任何产品来得到它。我认为关键不只是烘干机,还有你用的洗发水。我使用含有角蛋白的洗发水和离子超轻的护发素。底线-这吹风机是不可思议的!

15.>(5 星评级):我的头发很浓密,这确实减少了我干燥的时间。当我做完头发的时候,我用冷按钮,它似乎让我的头发更容易打理。

16.<最好的投资!>(五星评级):我现在可以通过少找我的发型师来省钱了。这台吹风机吹干我的头发比我以前用的吹风机快两倍。我真希望几年前就买了这个吹风机!

17.<优秀!>(五星级):我的客人浴室里有一台,客人可以使用。你不必去看厕所下面的人。

18.<T>(五星评价):这台烘干机太棒了,你的力量是首屈一指的。有点重,但值得它留下的结果。我很高兴。

19.>(5 星评级):它会把你家里的灯调暗,但是老天,它能把头发弄干!

20.我喜欢这个吹风机!>(五星评级):它能让我的头发很快变干,体积小小到可以装进我的行李箱里,它有一条可伸缩的线,让我的头发变得非常丝滑。而且价格也不能再低了。只要这款还在生产,我就不会再买别的牌子的。

21.<爱它!>(5 星评级):小,但有一些严重的热量和电力!

22.<爱它!>(五星评级):我可能在保留这台电吹风之前还买了另外 5 台。我和我的孩子们都有很多头发需要很长时间才能变干。这个吹风机是最好的;它比旧的干头发快得多。不用担心



不小心关闭或改变设置，你不必按住按钮进行冷拍，如果常规设置对你来说不够快，甚至还有一个涡轮增压设置。这节省了我们早上很多时间!我想再买一个放在另一间浴室里，这样我们俩就可以同时吹干头发了。刚开始的时候确实有点大，但没过多久就习惯了。其他烘干机的气流太大，我无法控制自己的发型，或者高温设置太热，但这台刚刚好。

23.这不仅是一个好产品，而且方便。>(5 星评级):正是我们想要的。

24.>(5 星评级):我买这个是因为它的好评。我染头发(掩盖灰色)，吹干头发总是花很长时间，头发看起来很干，到处都有一些翘起来。我买了这个，天哪!!它要花一半的时间才能变干，而且头发也不会飞起来!!我得到了很多赞美。我喜欢它，但从没想过会有人注意到。我真的得到了这么好的赞美。记得吗，我已经上了年纪，头发似乎越来越干了，染发对我没有任何帮助，除了掩盖灰色，我很高兴。但如果你觉得头发变得太干或太卷，头发会飞起来——这款护发工具就可以了。这是一次性购买，非常值得。3 个温度和一个我喜欢的酷按钮。我能说些什么呢，哦，它不是所有的沉重，常规的重量，除非你想用一个更便宜的吹风机吹一点和很多噪音。在那里做过…这是炸弹!!不要浪费更多的时间和金钱在令人沮丧的吹风机上，对自己好一点，买这个棒极了的吹风机。小心，你不会后悔一秒钟。

25.<作品太棒了!>(5 星评级):非常实惠，评论准确。它比我的旧的那台安静而且快。当你有 Conair 的品质时，没必要花很多钱买烘干机。

26.<好干燥!>(五星级):在纽约的一家酒店发现了这颗小宝石。干燥头发，不要留下卷曲或毛毛。可折叠手柄，方便存放和旅行。

27.<大吹风机!>(五星级):这是我用过的最轻、最安静的吹风机。事实上，我丈夫现在想要一个自己的!

28.<那么热。>(五星级):我非常非常喜欢这个电吹风……洗头 30 分钟，我就洗完了……

29.<爱它!>(五星级):以前我每天都害怕被吹干头发。它又长又厚，半个小时后还是干不了。这个吹风机最多 3-4 分钟就能把我的头发吹干。它完全改变了我的生活。我又买了一件现成的。你永远不知道吹风机能用多久。

30.<爱它!>(五星级):我用这台烘干机来代替我的旧的，哇，它真的很有效果。我的头发是如此的柔软和闪亮。我甚至可以用它来吹干我 1 岁女儿的头发，但我不能用我的旧的，因为即使在较低的设置下它也会变得太热。它干得快，不烧。绝对值得!

31.<超级小发电机!>(5 星评级):我已经有一个了，但是因为我经常旅行，我有一个在家里，一个在我的旅行拖车。一个小烘干机里有大量的热量和能量。

32.<IN LOVE>(五星级):我喜欢它

33.<爱它>(5 星评级):一年后还在使用…喜欢这种可伸缩的电线

34.<我爱它>(5 星评级):它比我的旧的更快，当它卷进去的时候不需要额外的空间，我觉得非常方便。

35.<爱它>(五星):每天使用

36.>(5 星评级):一切都是他们说的，占用很少的空间和工作

37.>(五星级):这是我用了多年的那台的替代品。

38.<伟大!>(五星评级):贵但物有所值。专业电吹风，专业效果。减少了我在浴室里擦干头发的时间

39.<爱它!>(五星评级):我有浓密的头发，这个宝宝让我的头发很快干，没有卷曲!爱它!

40.<低吹风机>(5 星评级):这款吹风机非常适合旅行。可折叠的把手，便于包装。拉线器很棒

41.它的工作原理就像烘干机……>(五星级):我的电吹风除了吹干外，没有别的用途。哈哈它简单，廉价，而且吹出很多热空气……
它很符合它的目的…

42.<爱它!>(五星评级):干得太快了，太棒了!我从法国带了一个我妈妈多年前买的但不幸的是，



由于电压差，我从来没有得到相同的电力。我订了这个，它完全一样(颜色除外)。把我平常的头发剪成两半。它会冒烟而且很热，所以要小心!
(我有一头松弛的黑发)

43.<工作很有魅力!>(五星级):我考虑过花一大笔钱去买一个高端的吹风机，但后来我意识到这是愚蠢的……而且我很便宜，所以我就买了这个，我真的很开心!我的头发很长(好像长到后背的中间)，而且很厚，这个吹风机不到 10 分钟就能吹干我的头发。我最后一台烘干机坏了。ER 以避免被我丈夫和盯着看讨厌构建他的眼睛因为我花这么长时间准备,从而隐瞒任何餐我们试图从他去吃吧,我就不打扰干燥而卷我的头发,然后就岩石自然来回但有了这个新的干燥机我可以做我的头发,仍然走出房子很快(不管怎样,相对)。
>我强烈推荐这款烘干机!没必要把辛苦赚来的钱浪费在一台高端烘干机上;这个就行了!

44.<看起来很棒>(5 星评级):不再需要回报

45.<棒极了的产品>(五星评级):我们有一个超过 18 年的旧产品，它终于坏了!新型号比以前的型号好!

46.<伟大的烘干机>(5 星评级):购买这个来代替被踩坏的(意外)。这是一种非常专业的品质，没有昂贵的价格标签。

47.<强力>(五星评级):快速干燥，有很多光泽-我第一次接触这个吹风机是在我住过的一家五星级酒店-它可以快速干燥，有很多光泽。

48.<很棒的烘干机>(五星评级):想要和 Y 餐厅墙上的一样的品牌和功率。
完美。这正是我的配偶想要的。

49.<很棒的烘干机>(五星级):我的头发用这个烘干机太柔软了，都不会乱动。还有，它没有我以前那台便宜的烘干机那么响。

50.<伟大的吹风机>(5 星评级):这个吹风机的功率很大。墙上的安装保持它的方式，但随时可获得的需要。这是为我们更换的吹风机，当另一个停止工作时，我们购买了同样的一个。

51.<伟大的产品!>(5 星评级):我在一本杂志上读到约翰·弗里达(John Frieda)的全卷吹风机，它比它的评论更有价值。干的快，叶子的头发柔软，没有卷曲，有很多身体。这是我第一次有了自己的卷发。我强烈推荐这种产品。没有它就活不下去。

52.伟大的>(5 星评级):这正是我想要的，非常轻的工作

53.<大>(五星评级):完全符合预期

54.不再有缠绕的绳索!>(五星级):这款电吹风有很大的功率和不同的温度设置。价格和质量都很好，是很重的。自动反冲绳是一个奖励!

55.<好干燥!>(五星级):我买这个电吹风来代替我以前的热风工具，它在使用多年后坏了。这个吹风机比我的旧的还要好!它更轻，仍然有同样的快速干燥能力的最后一个。

56.<美妙。潮湿的。负担得起的。几分钟内安装好。优秀>(五星评级):设置非常容易和快速。不需要工具。只需要几分钟。我放了一个酒吧的凳子坐下，并用一条毛巾盖住通风口。就像其他人说的那样，需要在新鲜空气孔上盖上毛巾，以保持蒸汽在里面。我不知道为什么人们抱怨通气孔。花 2 美元你可以为他们买一条毛巾，花 2000 美元你可以买一个不同风格的蒸汽桑拿。说真的，伙计们，买条毛巾吧。加一点香薰盐，这是幸福!仅用 200 美元就很难找到这样的幸福。我在 30 分钟内把我的体温烧到几乎让人难受，然后降低温度，进去，重新设置计时器。图片显示是一个小 off，因为门打开从一边，而不是中间的图片。不过不用担心。它很好。我会再买的。

57.<世界上最好的吹风机!>(五星级评级):太棒了 !!!!!!!!!!!!!!!!!!!!!

58.爱死这个吹风机了!>(五星级):这台吹风机很好——它很轻，而且吹干我头发的时间是我用过的所有其他吹风机的一半。

59.>(5 星评级):当我收到这款吹风机时，我真的很惊讶，因为产品的质量非常好。
这款吹风机非常便宜，外观和感觉都很棒，散发出很好的热量。我一定会把这个产品推荐给朋



友，我也很喜欢。

60.<我们爱它>(5 星评级):我把护发素添加到我的日常生活中。我丈夫发现，当他同时使用洗发水和护发素时，让护发素开着几分钟，效果很好

61.<爱我的吹风机>(五星级):没有它我就活不下去!

62.<最佳周边>(5 星评级):这是我购买的第三座墙架。我喜欢这种方便，而且不用放在抽屉里。

63.我这辈子最棒的电吹风!!>(五星评级):无。

64.>(五星级):我几个月前买了这个吹风机来替换我的旧的(不同品牌的)。这台新的吹风机和我的上一台一样强大(我很喜欢它)，但是安静多了，而且我的头发在吹风机之后也不觉得卷曲了。我唯一的抱怨是手柄底部有点脱皮。

65.<超级安静>(5 星评级):非常非常安静。它有最长的电线。这是我在烘干机上见过的。我不懂离子的事。

66.<太好了!>(五星评级):这把我的干燥时间缩短了一半!它比我预期的要重一些，但还是超出了我的预期。

67.<优秀!>(5 星评级):这个东西对我长满老茧的脚和我的角质层产生了奇迹。除了抗菌，你不能错过这个产品。只有很少的东西能持续下去。

68.<伟大的吹风机>(5 星评级):生活这个吹风机

69.<伟大的吹风机>(五星评级):这个吹风机真的能让你的头发更柔软，我已经使用它一个月了，我只有在我睡觉前或外出时头发是湿的情况下才使用它，它只需要 5 分钟就可以吹干我的头发(我的哥哥花了 30 分钟)，我高度强调它。

70.<一个伟大的干燥器。>(五星评级):当你不喜欢坐在吹风机或 blw 吹干你的头发时，这是你需要的。我可以带着它在房子里到处走，它会让你的头发变干。

71.完美>(五星级):我拿了一个我们国家的吹风机。但它不工作，因为。不同的电压。所以我不得不买一个新的，然后我选择了它，两天就买到了。太快了!

很好@!

72.完美>(五星级):正是我想要的。小,但强大的。

73.<伟大的产品!>(五星评级):这是一个伟大的烘干机!我的第一个冰箱在连续使用了 5 年后终于坏了，我也尝试了一个类似的金质微波炉，第一次使用时它就坏了，而且过热了。我很快就订购了 Andis，没有任何问题!我有一头长长的卷发，用不了 20 分钟就能把头发干透。我穿完刚买的这款产品以后，一定会再买第三次。

74.<好主意>(五星级):我们有一个小浴室，没有储物空间。这个电吹风太完美了。不使用的时候，它是不碍事的。不占用抽屉空间。甚至还有一盏夜灯。

75.<好产品>(五星级评级):干得快，让我的头发有光泽和直发。声音有点大，但是会放出很多空气，所以我想这是意料之中的。

76.<优秀产品>(五星评级):烘干机重量轻，散发大量热量。使头发迅速变干和光滑。将干燥时间缩短一半。不会再用标准烘干机了。

77.<了不起的!>(五星评级):即使是最薄的头发也会卷起来，而且卷发会持续一整天。瞬间发热，不会烧到头发。发胶和/或造型胶会延长卷发的寿命，但这不是必须的。我可以在 15- 20 分钟内自己做整个头部，而且我有长发。我发现的唯一不好的一面是，如果你烧伤了自己——你肯定会知道的。孩子们使用不安全，因为它很热。

78.<作品太棒了!>(5 星评级):快速烘干我的头发。我的头发很厚，这个吹风机把我的时间缩短了一半。我会向任何人推荐它。

79.<伟大的产品>(五星评级):我已经用这个完全一样的吹风机 20 年了!我一直买它是有原因



的，即使它在 4-5 年的持续使用后似乎会消失。这个夜灯非常适合主浴室。现在我们不用在半夜打开普通的灯来叫醒对方。

80.<神奇:>(五星级评级):我一直在使用这个洗面奶去年 3 年了,我所能说的就是这张脸洗奇迹,它清除所有我的痤疮,粉刺留下的所有我的伤疤,并使我的皮肤柔软,当然这可不是仅仅是因为洗面奶,我不得不穿上面部乳液为粉刺真的消失但是这个洗面奶也没有什么,不重皮肤上相比其他许多粉刺洗面奶我试过了

82.<棒烘干机>(五星评级):我已经连续两年想要这个作为圣诞礼物,终于我的家人买了它,现在他们想要一个!

83.最佳>(5 星评级):我有过的最好的吹风机。我的第一个笔记本每天使用了 12 到 15 年,已经记不清它的使用年限了。我猜它是在染发年龄 110 岁时死去的。

84.>(五星级):这台烘干机是我在未婚妻家过夜的时候买的,这样我不在的时候就可以把它放在他那里。它折成两半,绳子缩回手柄内,使它非常小。我立刻被这种力量所震撼,我的头发很浓密,但我可以在几分钟内完全吹干它。当我们结婚买了房子,它一直和我在一起,成为我的主要吹风机。我已经每天使用它大约 3 年了,它没有动摇,绝对值得!

85.<伟大的购买。>(五星级):虽然我买这是为了旅行,但我更喜欢它比我的普通大小的烘干机。它是粉红色的,但是在图片上看起来是紫色的,这就是我给它 4 而不是 5 的唯一原因。

86.<大吹风机!>(五星级):我从来没有意识到吹风机竟然可以让你的头发感觉柔软!这款能做到,2 秒钟就能吹干我的头发。也许不是 2 秒,但真的很快。

87.<快乐客户>(五星评价):达到我的期望。

88.<完美产品>(五星评级):产品出货及时,与广告完全一致。不要太多,也不要太少……洗完澡后帮我吹干头发和我的吉娃娃狗。

89.最佳>(五星评级):拥有吹风机几个月了…没有问题,不会过热…最好的投资

90.最佳>(5 星级):唯一一款两侧通风口的吹风机。这对我的工作是有必要的,更新的 backk vent 对我是没有价值的。

91.<爱它。>(五星评定):烘干机按时交货。它力量强大,2 分钟就能吹干我的短发,而且不重。爱死它了。

92.<我拥有过的最好的>(5 星评级):这个吹风机的优点我说不完。我有一头长长的卷发,花了一个多小时才吹干。现在我可以按时上班了。我以前的烘干机更贵,甚至没有接近好或快。没有理由再花更多的钱。我是一个非常满意的顾客。毫无疑问,下次我需要一个新的吹风机时,我会是一个老顾客。

93.>(五星级):独创产品。简单和方便。我把它粘在镜子上,待在那里。用一只手拿着烘干机刷洗也不麻烦

94.<五星>(五星评级):气压大

95.这是我用过的最好的吹风机!>(五星评级):快速吹干你的头发,不用油炸!!

96.<works great>(五星评级):它能做一个好的烘干机应该做的所有事情。不幸的是,可伸缩的电线并没有发挥应有的作用,但可以管理。

97.<爱它!!对我的购买感到非常高兴。>(五星级):我买了这个吹风机来代替我从朋友那里借的一个吹风机,那个吹风机因为我没有正确使用变频器而烧坏了。因为价格非常合理,再加上我得花 25 美元才能拿到免费运费,所以我买了两个,留了一个给自己。我每天都用这台烘干机,真是高兴极了。我的头发有点长(肩部以下),这个吹风机轻便有力,吹干我的头发很快。它可以做成很棒的旅行烘干机。它的效果比大多数酒店房间里的要好得多。以这个价格你不会出差错的。它有双重电压是一个很好的优点。

98.>(五星评级):喜欢这个电吹风。所描述的一切。说真的,大家按住插头,按下按钮收回,没有人会受伤!



99.<我爱它>(5 星评级):拥有我想要的一切, One hundred.<伟大的旅行吹风机。>(五星级):正好适合旅行的大小和功率。

101.这东西太棒了!>(五星级):当我偶然发现这个问题的时候,我在想为什么不找点别的东西呢?我对这次购买非常满意。另一位评论者说,我的头发很好,齐肩长发,而且没花一个小时就变干了。(当然,我也没有试图在头发湿漉漉的时候吹干头发。在吹干头发之前,我用毛巾把头发包起来,做了一会儿其他的事情。)不管怎么说,一开始我很担心,因为包装和大部分说明都是日文的。但我用的是低电压,还不错。我用普通的发刷来吹干头发,因为我还不知道怎么用圆形发刷。我吹干了大部分头发,然后用这个倒过来,让头发有一些提升和反弹,确实如此。而且,我的头发看起来不像使用普通吹风机时那样卷曲(我相信是因为有时普通吹风机既伤害头发又容易使头发打结)。我的头发实际上看起来既闪亮又光滑(就像我平头熨过一样)。如果你喜欢用吹风机吹干或熨烫头发,但看起来损伤小得多,我会尝试一下这个方法。我认为你可以只使用这台烘干机和一点精华液(取决于你的头发有多自然直发)。我很兴奋使用这个,并尝试找出圆刷。

102.<好产品>(五星评级):我妻子用过很多吹风机,我不敢相信她对这台吹风机有多满意……完全值得!

103.>(5 星评级):三种速度(噪音)和两种温度设置。我 4 岁的儿子不再抱怨吹头发会导致 queiter 速度。以这个价格买得好。

104.<伟大的价值!>(五星级评级):瞬间加热,头发干燥非常快,它比我使用过的任何其他品牌安静得多……很值这个价钱。

105.<完美>(5 星评级):做它需要做的事情。坚固耐用。变得如此火辣让你想要打你妈妈。买它。

106.<爱这个吹风机>(五星评级):这一个干得比我以前的吹风机快得多,似乎提供一个更平滑的完成。推荐。

107.最好的电吹风…句号!>(五星评级):真实故事…我花了大约 500 - 700 英镑在不同的头发和吹风机上。从便宜的到昂贵的我都有,这是到目前为止我拥有的最好的该死的吹风机。我是一名黑人女性,头发自然的 4c,这意味着厚和粗糙,通常吹干我的头发时,感觉很脆,但这台烘干机没有让我的头发感觉那样。当然,在吹干我的头发之前,我使用了 LOC 方法和 Alikay Naturals 产品,它变成了我需要的直发。我非常高兴和满足。我也用头发挑你吹干我的头发,它没有弹出,我也没有打破牙齿在挑。炸药!如果它确实弹开,使用耐热双面胶带保持它。

108.<爱这个电吹风>(五星级):我用这个电吹风好几年了,直到插头上的电涌保护器按钮不停地弹出。我在沃尔玛(Wal Mart)买了一个 Conair 吹风机来替换它,回到家后,插上电源,简直不敢相信它竟然这么便宜,这么响。我立刻把它退了回去,并且很高兴地发现 Remington 可以在亚马逊上买到。我完全被这个吹风机迷住了。电机安静,干燥时间短。唯一的问题是一开始感觉有点重。但是用了几次之后,我想我已经习惯了,因为现在这已经不是问题了。

109.<不要让便宜的价格愚弄了你>(5 星评级):我的电吹风坏了之后,我需要一个新的,我发现他们不再生产它了。我在当地的美容用品供应链上买了一个吹风机,以为它会和我那台已经去世的老古董一样好用。错了!它让我的头发感觉非常干燥,尽管我在吹干之前用过这种产品。我本打算投资一台 CHI,但在看到大多数关于那个型号的负面评论后,我决定省钱买一台 Andis 电吹风。我的头发又粗又长(超过我的肩膀 4 英寸),现在我的头发越来越松弛,越来越自然。我有大约 3-4 英寸的新生长,其余的仍然是放松的,所以直发样式暂时工作。这个吹风机就是一切!它不仅节省了钱包,而且还能让我的头发很快变干,同时不会吸干头发上的水分。吹干我的头发不到 20 分钟,甚至比我原来的吹风机还要快,而且把我的发根弄得很直(我用的是小梳子)。它很容易握,而且很热,不会烫到头皮。加上我的 Sedu



扁铁，我的头发几乎是沙龙的质量，超级柔软和闪亮!我会向任何和我有相同发质的人推荐这个。我非常高兴!

110.伟大的吹风机>(5 星评级):我买这个来代替另一个雷明顿我想

用了几年。到目前为止，我对我买的东西很满意。我选择这个是因为我过去吃过的 Remingtons 的质量，而且这个看起来和其他的都一样。质量。吹风机确实有轻微的橡胶感觉。我从来没有吃过这样的。不能说我在乎;这只是不同。没有好坏之分。

我特意选择了这一个，因为附加了扩散器。对于我们这些卷发的人来说，这是必不可少的，而且效果很好。

111.<大干燥机...>(5 星评级):暂时拥有它。到目前为止没有问题。和我第一天买的时候一样好用。一些不错的功能为钱

112.<honest>(5 星评级):我和妻子每天都用这个产品至少 10 年了，最后它终于坏了，所以我不得不买一个新的。

113.伟大的电吹风包装成小尺寸!>(五星级):完美尺寸。做我想做的事!!不抱怨!

114.<好产品>(五星评级):我有一个 Andis 吹风机至少 8 年了，直到最近它在我身上死去。
替换太好了;产品还是像以前一样可靠!

115.<很棒的烘干机>(五星评级):我用的很好，因为我每天都用。此产品无投诉。易于使用和处理。

附录 D VADER 情绪分析的源代码

```
from vaderSentiment.vaderSentiment import SentimentIntensityAnalyzer
#VADER is freely available at https://github.com/cjhutto/vaderSentiment
data = pd.read_csv(loadurl, sep='\t') [['star_rating', 'review_headline', 'review_body']]

#VADER Model
def transform(df):
    d=df.values.tolist()
    for element in d:
        vs1 = analyzer.polarity_scores(str(element[1]))
        l1=list(vs1.values())
        vs2 = analyzer.polarity_scores(str(element[2]))
        l2=list(vs2.values())
        element.append(l1,l2)
    return pd.DataFrame(d)
data=transform(data)
data.columns=['star', 'head', 'body', 'h_sn', 'h_wn', 'h_m', 'h_wp', 'h_sp', 'b_sn', 'b_wn',
              'b_m', 'b_wp', 'b_sp']
data.to_csv(outputurl)
```

附录 E 信息性算法的源代码

```
def cosdistance(v1,v2):
    vec1=v1
    vec2=v2
    fenmul=1
    fenmu2=math.sqrt(vec2[0]*vec2[0]+vec2[1]*vec2[1]+vec2[2]*vec2[2]+vec2[3]*vec2[3]+vec2[4]*vec2[4])
    fenzi=vec1[0]*vec2[0]+vec1[1]*vec2[1]+vec1[2]*vec2[2]+vec1[3]*vec2[3]+vec1[4]*vec2[4]
    return 1-fenzi/(fenmul*fenmu2)

def entropy(v1):
    total=0
    vec=v1
    for i in range(5):
        total+= vec[i]*math.log(vec[i])
    return total

#change data to one-hot-vector form
def onehotvec(df):
    data=df.values.tolist()
    result=np.zeros([len(data),5]).tolist()
    for i in range(len(data)):
```



```
#data import
data=pd.read_csv(loadurl)
intvec=onehotvec(data['star']).values.tolist()
vecvec=data[['p1','p2','p3','p4','p5']].values.tolist()
help=pd.read_csv('D:\Problem_C_Data\hair_dryer.tsv',sep='\t')['helpful_votes'].values.tolist()
date=pd.read_csv('D:\Problem_C_Data\hair_dryer.tsv',sep='\t')['review_date']
#get informative value
cal=[]
for i in range(len(help)):
    x=(1+help[i])*pow(np.e,(-1)*cosdistance(intvec[i],vecvec[i]))*pow(np.e,1+entropy(vecvec[i]))
    cal.append(x)
result=pd.DataFrame(cal)
```

附录 F 名誉计算的源代码



```
def sigmoid(x):
    return 1/(1+pow(np.e,(-1)*x))
def get_reputation(df):
    data=df.values.tolist()
    data=data[::-1]
    curve=np.zeros(len(data)).tolist()
    curve[0]=1
    #param set
    p=0.75
    n=0.25
    k1=25
    k2=20
    #k=2
    for i in range(1,len(data)):
        if data[i][0]>=0:
            if data[i][3]==0:
                curve[i]=curve[i-1]+(p-n)*data[i][0]
            else:
                curve[i]=curve[i-1]+(p-n)*data[i][0]-sigmoid(k1*curve[i-1])*k2*data[i][1]/data[i][3]
        else:
            if data[i][3]==0:
                curve[i]=curve[i-1]+(p+n)*data[i][0]
            else:
                curve[i]=curve[i-1]+(p+n)*data[i][0]-sigmoid(k1*curve[i-1])*k2*data[i][2]/data[i][3]
    return pd.DataFrame(curve)
#calculate
curve=get_reputation(inputdata)
```

附录 G Beyes 模型的源代码

```
def word_freq(df,s):
    data=df['body'].values.tolist()
    count=0
    for i in range(len(data)):
        if s in data[i]:
            count+=1
    return count/len(data)
def countfive_prob(data):
    d=data['star'].values.tolist()
    total=len(d)
    count=0
    for i in range(len(d)):
        if d[i]==5:
            count+=1
    return count/total
def countfive_word(df,s):
    count=0
    total=0
    data=df[['star','body']].values.tolist()
```



关注科研交流



```
for i in range(len(data)):
    if data[i][0]==5:
        total+=1
        if s in str(data[i][1]):
            count+=1
return count/total
def train(str):
    s=str
    data=pd.concat([data1,data2,data3],axis=0)
    data=data[['head','body','0','star']]
    five_prob=countfive_prob(data)
    word_star=countfive_word(data,s)
    word_frequency=word_freq(data,s)
    result=word_star*five_prob/word_frequency
    return result

textdata=pd.read_csv(loadurl1)[['review_body','review_head']]
wordlist=pd.read_csv(loadurl2)
#get beyes prob value of each highly frequent word
result=[]
for i in word:
    result.append([i,train(i)])
r=pd.DataFrame(result)
```

附录 H 时间序列预测的源代码

```
from statsmodels.tsa.ar_model import AR
# train autoregression
model = AR(train)
model_fit = model.fit()
window = model_fit.k_ar
coef = model_fit.params
# walk forward over time steps in test
history = train[len(train)-window:]
history = [history[i] for i in range(len(history))]
predictions = list()
for t in range(len(test)):
    length = len(history)
    lag = [history[i] for i in range(length-window,length)]
    yhat = coef[0]
    for d in range(window):
        yhat += coef[d+1] * lag[length-d-1]
    predictions.append(yhat)
error = mean_squared_error(test, predictions)
print('Test MSE: %.3f' % error)
print('Test RMSE: %.3f' % math.sqrt(error))
```

附录 I Wordcloud 图片的源代码

```
from wordcloud import WordCloud, ImageColorGenerator, STOPWORDS
txt=pd.read_csv(loadurl1)
my_wordcloud = WordCloud(
    background_color='white',
    mask=imagedata.imread(pictureurl), #param set
    max_words=100,
    max_font_size=100,
    random_state=50,
    width=1000,
    min_font_size=20,
    relative_scaling=0.4,
    color_func=genclr,
).generate(txt)
# show word cloud picture
plt.imshow(my_wordcloud)
```

