通过先进的分析见解增强用餐体验：案例研究
饼干之爱：峡谷

执行摘要

我们的项目利用先进的数据筛选和文本挖掘方法来分析 Biscuit Love: Gulch（位于纳什维尔的一家著名早餐和早午餐餐厅）的业务绩效。通过利用 Yelp 的综合数据集（其中包括 2012 年至 2022 年的 4,247 条评论），该分析旨在深入研究客户反馈的动态，以挖掘可操作的见解，从而使商家和消费者都受益匪浅。

项目目标

在数字时代，用户生成的内容已成为品牌与消费者之间的重要桥梁。评论和评级不仅影响潜在客户的购买决策，还为公司提供有价值的反馈，帮助他们改进产品和服务。然而，传统的星级评定系统往往无法充分捕捉消费者的复杂情感和具体意见。因此，我们采用了一种新的分析方法，通过情感分析来细分正面和负面评论中的主要话题，旨在揭示消费者的真实感受和更深层次的需求。

从平台的角度来看，目标是通过更深入地了解商家的运营优势和需要改进的领域来增强商家的能力。从消费者的角度来看，目的是帮助顾客选择更符合他们的喜好和期望的餐饮场所。因此，我们以 Biscuit Love: Gulch 餐厅为例进行分析。

他们可以通过以下方式利用我们的分析结果来提高他们的吸引力和声誉： 1.分析顾客评论，以确定餐厅做得好的地方以及需要改进的地方。

2.实施基于情感分析的新评分系统，提供更客观的评分，帮助餐厅提升服务。

数据说明

Yelp数据集包含2012年至2022年美国不同餐厅的历史评论。该数据集包括商业信息文档、评论文档和客户ID文档，但我们只关注商业信息和评论信息。整个评论数据集近700万行9列，包括评论ID、用户ID、商家ID、评分星级（范围为1到5）、评论和评论日期。我们关注前 20 家商店，总共有 96000 行和 10 列。由于评论数据集仅包含商店 ID，不包含商店名称，因此我们从商业信息数据集中提取商店名称。我们分析的目标之一，Biscuit love: gulch，在数据集中有超过 4000 条历史评论，是纳什维尔排名前 20 的餐厅之一。

然而，也存在一些缺点：

1. 评论包括多种语言。

2. 有大量拼写错误的单词、流行语和表情符号。
3. 评论中夸大其词。

方法

关键词提取

这部分主要是根据我们收到的评论来识别前 10 个最积极和最消极的关键词。我们实现两种方法。

具有 TF-IDF 的 N 元语法

步骤1.我们根据星级列将企业的整体评论分为两组，即好评（4、5 星）和差评（1-3 星）。

步骤2.我们使用TF-IDF将文本格式评论向量化为数字矩阵，并将n-gram范围设置为1-3，这意味着我们将考虑一个单词一元组到三个单词三元组之间的短语单词，和企业主更容易理解上下文。

步骤3.计算每个短语的平均TF-IDF得分，然后根据每个短语的平均得分，通过对两组数据进行降序排序，选出最积极的10个短语和最消极的10个短语。

套索回归

步骤 1. 与上一步类似，我们使用 n-gram 将整个评论拆分为一到三个单词的短语，并使用 TD-IDF 矢量器将字符串格式变量转换为浮点矩阵。

步骤2.将矩阵放入Lasso模型中。由于我们想要探索星星与每个短语之间的关系，使用短语来预测星星，因此我们将这两个变量放入模型中。

a) 为什么选择套索：
   i. Lasso回归可以进行特征选择，它可以将系数缩小到零以从模型中删除不太重要的短语。我们的短语栏太多了，这个功能对我们来说是有好处的。
   ii. Lasso回归在损失函数中加入L1惩罚，可以避免过拟合问题。步骤 3.结果回顾，Lasso回归的输出中我们主要关注两个变量，即短语及其相应的相关系数。从相关系数中，我们可以了解短语和星号之间的强度和方向。如果相关系数趋于较大的正数，则表明短语之间的相关性较强

积极评论和高星级，而负面评论的数量往往较小，因此该短语与负面评论和低星级有很强的相关性。

步骤4.结果输出：按照相关系数值对短语进行排序。 Top 10 积极短语：系数最大的短语。前 10 个负面短语：系数最低的短语。

情绪分析

情感分析实现了VADER词典，该词典在处理社交媒体内容方面是先进的，它会考虑表情符号、标点符号、俚语、缩写词等互联网上常用的表达方式。

步骤1.对评论进行初步情感分析。我们将复合极性设置为 1-5 五组情绪（1：非常消极，2：消极，3：中性，4：积极，5：非常积极）。

步骤2.根据初步的情感分析结果，得到最高分和最低分

主要在0.9到-0.9之间，因此我们将每组之间的距离设置为0.36（1.8 / 5）。因此，我们设置的阈值是[-0.54,-0.18,0.18,0.54]。

步骤 3. 在原始数据框中创建一个新列，并将新的情绪分数存储到该列中。除此之外，平均分数是通过结合我们拥有的星级和我们生成的情绪类别来计算的，从而提供了对整体情绪的更客观的看法。

用户界面（图1）

我们使用ipywidgets包让用户可以切换他们想要分析的业务，然后选择他们想要执行的分析。

步骤1.选择企业：从数据框中获取唯一的企业名称到列表中，然后用户可以从列表中选择企业。

步骤2.选择任务：选择“情感分析”和“关键词提取”，对已选择的业务进行分析。

结果

我们以《饼干之恋：峡谷》为例来调整和展示我们的分析和方法。所以，结果我们也会重点讲这个业务结果。

关键词提取：

从关键词提取结果中，我们得到这些可能有意义且有趣的短语，这可能可以获得一些商业洞察。从积极的方面来看，我们了解到顾客喜欢讨厌的公主、奖金、慢性培根（商店里的三餐名称）、热鸡和橙汁。而它的早餐时段是全天最受欢迎的时段。从不好的地方，企业主可以了解到自己还有哪些地方需要改进，像有的顾客说“饼干干”、“饼干硬”，老板和老板可以想想在购买前是否可以询问一下顾客的口味，制作它（类似于牛排馆）。另外，也有人说“太甜了”，所以业主需要考虑减少糖的添加。（图2、3）

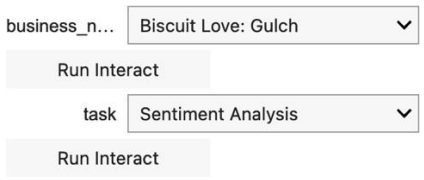
情绪分析：

从结果表中我们发现，即使评价很正面，顾客也倾向于给予较低的星级，但这符合顾客的习惯，消费者会比较严格，他们可能认为它并不完美。然而，没有一家商店是完美的。（图4、5、6）

结论

我们使用文本挖掘和情感分析对 Biscuit Love: Gulch 的 Yelp 评论进行分析，为客户体验和偏好提供了宝贵的见解。 TF-IDF 和 Lasso 回归等技术确定了优势和需要改进的领域，例如独特菜肴的受欢迎程度以及食物质地和甜度的问题。这些发现提出了提高服务质量和客户满意度的可行步骤。此外，我们开发的用户友好界面可以轻松探索数据，帮助企业主和客户做出明智的决策。该项目强调了数据驱动方法在理解和改善竞争激烈的餐饮业中的就餐体验方面的重要性。

人物



（图1：交互界面）

ngram result:
Top Bigrams for Good Reviews (4 and 5 stars):

	ngram	tfidf_score
13966	comparably star granted	0.023016
65994	din time	0.016137
118448	hot mess chicken	0.014659
110656	place satisfy instagram	0.014587
19228	orange juice bit	0.014422
91353	friday afternoon take	0.014062
27614	comfortable eat want	0.013162
102847	place definitely hype	0.012825
39803	hint	0.011610
104249	provide east	0.011072

Top Bigrams for Bad Reviews (1 and 2 stars):

	ngram	tfidf_score
5301	brunch pancake pantry	0.032386
26958	bar co	0.022062
49075	biscuit typical trendy	0.018731
45660	average review thing	0.016503
11353	huge portion tho	0.015007
28275	heavy sweet heavy	0.013356
37745	land nashville	0.012152
8047	good quite	0.011788
53116	everyone nice east	0.010439
36964	real frankly	0.010223

****

（图2：TF-IDF与N-grams的输出结果）

****

Lasso Result:
Top 10 Positive Phrases:

	phrase	lasso_coeff
52035	nasty princess	1.758564
6690	best breakfast	1.686129
50961	must nashville	1.110952
37718	hot chicken	0.969071
22896	east nasty	0.895007
9418	bonuts must	0.784087
13828	chronic bacon	0.653324
25872	everything delicious	0.538861
55174	orange juice	0.427146
6684	best biscuit	0.304508

Top 10 Negative Phrases:

	phrase	lasso_coeff
32623	good biscuit	-0.905967
50537	much good	-0.969096
32283	give star	-1.237352
7560	biscuit good	-1.459053
7266	biscuit biscuit	-1.809805
7581	biscuit hard	-1.856910
81510	tourist trap	-2.074740
53587	nothing special	-3.352368
7435	biscuit dry	-4.423304
72198	somewhere else	-4.618242

（图3：N-grams与Lasso的输出结果）

	VADER_polarity	VADER_score	text	stars	avg_stars
91793	5	0.5859	bonuts yogurt amazing seat arrive immediately ...	5.0	5.0
91794	5	0.9274	friendly service delicious cute atmosphere bon...	5.0	5.0
91795	5	0.9849	love definitely right word place sunday mornin...	4.0	4.5
91796	5	0.9842	nashville first time bff biscuit love definite...	5.0	5.0
91797	5	0.9805	biscuit love little spot tuck minute downtown ...	3.0	4.0
...	...	...	...	...	...
96035	5	0.9954	wow trip nashville highlight trip biscuit love...	5.0	5.0
96036	5	0.9595	yummmm biscuit arrived around 30am friday wrap...	4.0	4.5
96037	5	0.9313	good long eventually super nasty princess spic...	5.0	5.0
96038	5	0.9473	always little wary place hype long overall exp...	4.0	4.5
96039	4	0.4767	nice place eat pack may want early sunday	4.0	4.0

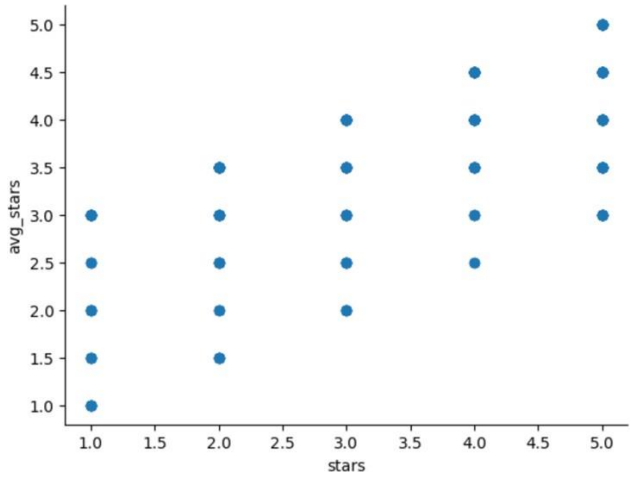
4247 rows × 5 columns

（图4：VADER情感分析输出结果）

	VADER_polarity	VADER_score	text	stars	avg_stars
0	5	0.8555	twice brunch enjoy immensely everything delici...	4.0	4.5
1	5	0.9371	first meal new orleans lunch special seafood s...	4.0	4.5
2	5	0.9931	service excellent atmosphere raw bar fantastic...	4.0	4.5
4	5	0.9589	oyster happy hour 7pm boyfriend share dozen oy...	4.0	4.5
5	2	-0.1949	place suck terrible service overprice mediocre...	1.0	1.5
...	...	...	...	...	...
4649	5	0.9413	staff super friendly attentive oyster good hap...	4.0	4.5
4652	3	-0.0133	wife across restaurant accident happy happen r...	5.0	4.0
4654	5	0.9871	guest hilton restaurant locate onsite first im...	3.0	4.0
4655	5	0.9459	little weird review john besh restaurant sexua...	4.0	4.5
4660	5	0.9524	definitely musteats best shrimp grit crab au g...	4.0	4.5

2264 rows × 5 columns

（图5：结果只保留预测星与我们已知星值不同的记录）



（图6：我们计算的平均星数与我们已知的星数之间的关系）

引文

数据来源：Yelp 开放数据集。 Yelp 数据集。 （日期不详）。 https://www.yelp.com/dataset