**Advances in Applied Mathematics** 应用数学进展, 2022, 11(5), 3078-3084
Published Online May 2022 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2022.115328

Hans汉斯

# 网红营销基于情感分析的消费评论数据的挖掘

**王　哲，杜凤娇**

徐州工程学院数学与统计学院，江苏　徐州

收稿日期：2022年4月27日；录用日期：2022年5月21日；发布日期：2022年5月31日

## 摘　要

**随着2020年疫情爆发对经济的影响，常规的商业营销模式受到了冲击。以网红营销为代表的电子商务迅速崛起。电商平台作为网红营销的主要渠道，成为了消费者挑选、评价电商产品的媒介。因此，在电商平台的消费者评论具有文本挖掘的价值和潜力。本文使用网络爬虫技术，爬取到49,700条天猫商城评论数据其中有效数据为40,138条，采用TF-IDF算法提取评论中的关键词，通过Basic LDA模型对商品评论进行属性抽取，并运用变分贝叶斯推断对模型进行求解，最后通过朴素贝叶斯对评论进行情感极性分析，最终得到结论：消费者对电商产品的评论从功能性、消费理性两个角度出发，主题1、3、4表现出消费者重视电商产品使用的直观感受的倾向，主题2表现出消费者重视电商产品实际解决需求的功能性的倾向，消费者对电商产品的情感得分里积极、中性、消极分别为36%、63%、1%，总体上对电商产品持认可态度。**

## 关键词

**网红营销，LDA模型，变分贝叶斯推断，朴素贝叶斯**

# Influencer Marketing Based on Consumer Comment Data Mining of Sentiment Analysis

**Zhe Wang, Fengjiao Du**

School of Mathematics and Statistics, Xuzhou University of Technology, Xuzhou Jiangsu

Received: Apr. 27th, 2022; accepted: May 21st, 2022; published: May 31st, 2022

## Abstract

**With the economic impact of COVID-19 since 2020, conventional business marketing models have been hit. E-commerce, represented by influencer marketing, has risen rapidly. As the main chan-**

文章引用: 王哲, 杜凤娇. 网红营销基于情感分析的消费评论数据的挖掘[J]. 应用数学进展, 2022, 11(5): 3078-3084.
DOI: 10.12677/aam.2022.115328

**nel of influencer marketing, e-commerce platform has become a medium for consumers to select and evaluate e-commerce products. Therefore, consumer evaluation in e-commerce platform has the value and potential of text mining. In this paper, the web crawler technology is used to crawl 49,700 tmall review data, 40,138 of which are valid data. TF-IDF algorithm is used to extract key words in the review, Basic LDA model is used to extract attributes of the product review, and variational Bayesian inference is used to solve the model. Finally, naive Bayes was used to analyze the emotional polarity of the comments, and the final conclusion was drawn: Consumers' comments on e-commerce products are from the perspectives of functionality and consumption rationality. Theme 1, 3 and 4 show consumers' tendency to attach importance to the intuitive feeling of using e-commerce products, while theme 2 shows consumers' tendency to attach importance to the functionality of e-commerce products to solve actual needs. The positive, neutral and negative sentiment scores of consumers to e-commerce products are 36%, 63% and 1%, respectively. In general, consumers approve of e-commerce products.**

## Keywords

**Influencer Marketing, Latent Dirichlet Allocation Model, Variational Bayesian Inference, Naive Bayes**

Copyright © 2022 by author(s) and Hans Publishers Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/
Open Access

## 1. 引言

随着 2020 年疫情爆发对经济的影响，常规的商业营销模式受到了冲击。以网红营销为代表的电子商务迅速崛起。电商平台作为网红营销的主要渠道，提高了商业效率，同时降低了货物流通环节和仓储管理费用，降低了交易成本[1]。

电商平台鼓励消费者发表评论以提升服务质量并吸引更多消费。由于消费者在电商平台上对产品各个属性发表评论，这些包含购物和产品体验的信息成为了一种重要的商业资源，同时也成为了消费者挑选、评价目标电商产品的引导。因此消费者可以通过他人的评论了解目标商品各项属性质量是否符合自己的要求或优缺点，从而尽可能实现理性、科学的消费[2] [3]。对于商家来说，消费者评论是最直接有效的市场反馈信息，通过评论数据可以及时改进购物服务，提升产品质量，从而实现精准营销。

对评论的文本数据进行分析常常面临另一个难题，就是将文本这种非结构数据转化为向量类型的结构化数据[4]。常见的方法有 One-Hot [4]、TF-IDF [5] [6]、Word2Vec [7]等方法。将“评论文本”转换为数值型结构化数据是评论数据挖掘的核心之一。但是“One-Hot”方法只能反映出词语出现的频数，而不能反映词语之间的关系。“TF-IDF”则是“One-Hot”的一种改进，该方法将每个文档中出现的词向量除以总频数，可以同时反映词语频率和词语在文档中的重要程度。

在文档主题提取方面，M. Blei 在 2003 年提出 Basic LDA [8]，希望通过困惑度与主题数量的曲线来寻找出最高频且主题之间差异最明显的极值点，满足这个要求的点，曲线的纵轴越高，则该值越适合作为 topic 的个数。Yao [9] (面向微博)选择使用词汇之外的一些特征信息来判别文本的主客观分类，如标点符号、人称代词、数词等角度。

## 2. 评论数据预处理及词频分析

网红营销模式涉及到的产品种类多种多样，本团队选取具有代表性的食品、服饰、美妆、电子产品、

家居用品这五类作为研究对象，并选取各类商品评论在 2000 条以上的 5 个产品进行数据采集。

针对评论采集，本文主要使用了 Python 爬虫技术，基于 request 库的浏览器模拟访问数据采集技术，分析数据、网页结构以及网页数据流的形式和来源，编写程序后运行，抓取到天猫平台 25 个商品的 49,700 条数据。本文针对数据中残缺的、异常的、不一致的数据进行数据清洗。经过预处理后，得到有效数据 40,138 条。文本有效率为 80.7%。其中，大部分无效数据属于重复数据和自动评论。

本节为了对评论数据做进一步的分析建模，使用 Python 软件的 jieba 库完成中文的分词工作。同时，无实际意义的词汇会增加程序的运算成本，例如，“了”、“呢”、“啊”以及一些特殊符号和中文标点等等。将样本数据进行中文分词之后，再对高频词进行统计，并画出词云统计图如下所示：


**Figure 1.** Word cloud figure
**图 1** 词云统计图

图 1 中，字号与词频大小成正比。分析上图可知，通过天猫平台消费的客户基本抱有积极的反馈，如“不错”、“买”、“喜欢”等等。消费者也关注商品的性价比，如“效果”、“价格”、“味道”等等。商品的品牌及产品形象也是消费者的关注点之一，如“包装”、“外观”、“物流”等等。网红在销售的过程中显然也占据了很大的地位，如“直播间”“推荐”等词汇。

## 3. 朴素贝叶斯情感极性分析

本文使用 Python 中用于自然语言处理的 snownlp 库对评论数据进行情感倾向分析。通过调用 snownlp 中的贝叶斯分类器[10]生成评论数据属于积极情绪的概率，并依据小概率原理对数据进行情绪极性判别。

朴素贝叶斯(Naive Bayes Classifier)，假设文本数据中各类词汇之间相互独立，利用贝叶斯公式得到给定样本情况下样本属于某一类别的后验概率。具体公式如下所示：

$$\hat{y} = \arg\max_{y} \frac{P(y)\prod_{j=1}^{n} P\left(x_j \middle| y\right)}{P(x)},$$

上式中，$y$ 表示文本数据属于哪一种情感极性的标签值，即情感极性值，$x_j$ 表示总量为 n 的数据集中

的第 $j$ 条评论，$\hat{y}$ 表示待估参数，即需要进行判别的情感极性值大小，$P(y)$ 表示在所有评论数据中出现某种情绪(例如积极情绪)的概率大小，$P(x_j|y)$ 表示在带有某种情绪的所有数据集合中出现第 j 条评论的概率大小，$P(x)$ 表示抽取样本数据的概率，这一概率通常在给定抽样方法时就以确定，即抽样概率。

通过该公式可以求得后验概率 $P(y|x)$ 最大时，$y$ 的情感极性值，从而以生成概率的方式达到分类效果。判别公式如下：

$$y(x_j)=\begin{cases}1, & 0.95<p\le 1,\\ 0, & 0.05<p\le 0.95,\\ -1, & 0<p\le 0.05,\end{cases}$$

上式中 $y(x_j)$ 表示第 $j$ 条数据 $x_j$ 的情感极性值，$p$ 表示在给定第 $j$ 条数据 $x_j$ 时情感极性值 $y_j$ 等于 1 的条件概率，即该评论内容为积极情绪的概率。小概率原理是指小概率事件在一次实验中不可能发生。当评论数据为积极情绪的概率大于 0.95 时，认为该评论为积极情绪，小于 0.05 时为消极情绪，在 0.05 和 0.95 之间时，认为是中性情绪。分类的结果如下图：

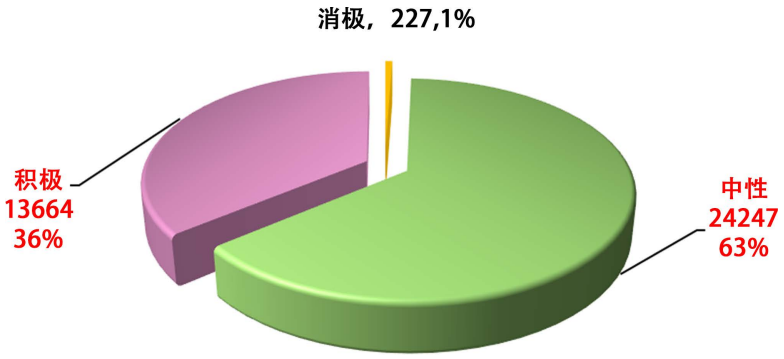


**Figure 2.** Sentiment analysis results
**图 2.** 情感分析结果展示图

根据图 2 可知，大部分消费者对商品的评价为中性和积极，网红推销的产品效果具有一定的真实性，可以在具体的调查分析中，针对产品的预期效果和实际效果进行问卷调查。

## 4. LDA 主题模型

为了获得消费者对于评价电商产品性能的侧重点，本文采取 Basic LDA 模型(Basic Latent Dirichlet Allocation)针对分词后的评论数据进行评论属性维度的抽取[11]。LDA 模型假设每篇文章为词袋模型，即忽视单词之间的顺序的影响，仅考虑单词在文章中出现的频率。因此 LDA 模型适合通过对名词进行主题分析。常用的 smoothed LDA 模型在求解时常使用 Gibbs 采样对隐变量的后验分布进行随机近似推断。Gibbs 采样具有推导简洁、易于理解的优点[12] [13]，但是求出的结果是随机解，输入参数都相同的情况下，每次求解结果都会改变，如果人为选择一个解释性强的结果，这样的筛选过程会导致结果产生偏差。因此，本文采取变分贝叶斯推断来对 baisc LDA 模型隐变量的后验分布进行确定近似推断[10]。Basic LDA 模型具体如下：

$$P(D|\boldsymbol{\alpha},\boldsymbol{\beta})=\prod_{d=1}^{M}\int p(\boldsymbol{\theta}_d|\boldsymbol{\alpha})\left(\prod_{n=1}^{N}\sum_{z_{d,n}} p(z_n|\boldsymbol{\theta}_d)\,p(\boldsymbol{w}_{d,n}|z_{d,n},\boldsymbol{\beta})\right)d\theta_d$$

$$\theta \sim Dir\left(\boldsymbol{\theta}\middle|\boldsymbol{\alpha}\right)=\frac{1}{\Delta\left(\boldsymbol{\alpha}\right)}\prod_{k=1}^{V}\theta_k^{\alpha_k-1},$$

$$\Delta\left(\boldsymbol{\alpha}\right)=\frac{\prod_{i=1}^{k}\Gamma\left(\alpha_i\right)}{\Gamma\left(\sum_{i=1}^{k}\alpha_i\right)},$$

$$w_{d,n}\sim Multinomial\left(\boldsymbol{\beta}\right)=p\left(w_{d,n}\middle|\boldsymbol{z}_n,\boldsymbol{\beta}\right)=\prod_{i=1}^{K}\prod_{j=1}^{V}\left(\beta_{ij}\right)^{\boldsymbol{w}_n^j\boldsymbol{z}_n^i}$$

上式中，$D$ 代表整个评论集数据，M 表示评论集中评论篇数，$\boldsymbol{\alpha}$ 为 $\boldsymbol{\theta}$ 的超参数 $\boldsymbol{\alpha}=\left(\alpha_1,\alpha_2,\cdots,\alpha_V\right)$，$z_{d,n}$ 指第 $d$ 篇文章中的第 $n$ 个词的主题分配，$\boldsymbol{\beta}$ 表示从主题 $\boldsymbol{z}^i$ 到单词 $\boldsymbol{w}^j$ 的产生概率(此处 i、j 为上标，$\boldsymbol{z}^i$、$\boldsymbol{w}^j$ 表示当前分配的主题是否为 $i$ 号主题，若是则为 1，否则为 0，$\boldsymbol{w}^j$ 也类似表示单词是否为 $j$ 号单词，若是则为 1，否则为 0)，即 $\beta_{ij}=P\left(\boldsymbol{w}^j=1\middle|\boldsymbol{z}^i=1\right)$，$\boldsymbol{\theta}$ 为该文档的主题分布，$\boldsymbol{\theta}=\left(\theta_1,\theta_2,\cdots,\boldsymbol{\theta}_d,\cdots,\theta_K\right)$，$w_{d,n}$ 是第 $d$ 篇文章的第 $n$ 个单词，服从参数为 $\boldsymbol{\beta}$ 的多项分布。

根据数据预处理阶段对评论数据分词的结果，使用 Python 中的 sklearn 包的 LDA 进行计算，本文设定模型每篇文章的超参数为 $\boldsymbol{\alpha}=\left(0.1,0.1,\cdots,0.1\right)_{1\times4}$，设定模型每个主题的超参数为 $\boldsymbol{\beta}=\left(0.01,0.01,\cdots,0.01\right)_{1\times N}$。软件对各主题下出现的评论词语计算并排序的结果如下：

**Table 1.** Table of LDA model's results
**表 1.** LDA 模型结果表

| 主题 1 | 主题 2 | 主题 3 | 主题 4 |
|---|---|---|---|
| 直播间 | 味道 | 效果 | 评论 |
| 价格 | 音质 | 感觉 | 用户 |
| 客服 | 外观 | 质量 | 电视 |
| 质量 | 口感 | 颜色 | 很漂亮 |
| 物流 | 电脑 | 评价 | 活动 |

由表 1 中主题分析的结果可以看出，主题 1、主题 3、主题 4 都倾向于对电商产品的直观感受和情感评价，例如主题 1 中的效果、味道、感觉，主题 3 中的口感、面料等词语。主题 2 则倾向于对电商产品的质量、售后服务、物流等角度进行评价，更偏重于产品及其服务如何解决用户需求的功能性问题。因此，可以将主题 1、2、3、4 分别命名为营销服务感知、商品感知、功能感知、众评感知。

根据 PyLDAvis 结果中对各主题之间的 JSD 距离数据使用主成分分析将 LDA 的分类结果降维至两个主成分，从而得到各主题在二维平面上的坐标。各主题的条件分布气泡坐标如图 3。

根据上文对各主题含义的分析可知，主题 1、2、3、4 分别代表营销服务感知、商品感知、功能感知、众评感知。首先对第二主成分(PC1)进行分析，从左至右主题分别为：众评感知、商品感知、营销服务感知、功能感知。从该排序可知，用户群体的主题逐渐从网红产品的宣传特点转向实际的功能特点，并且对网红产品的实际满足消费者自身需求的感知不断增加。因此，可以将第一主成分命名为功能性维度。对第二主成分(PC2)的含义进行分析，从上至下主题的排序分别为：众评感知、营销服务感知、功能感知、商品感知。从该排序可以看出，消费者对于网红产品的评论主题从他人的评价、商家对产品的评价或宣

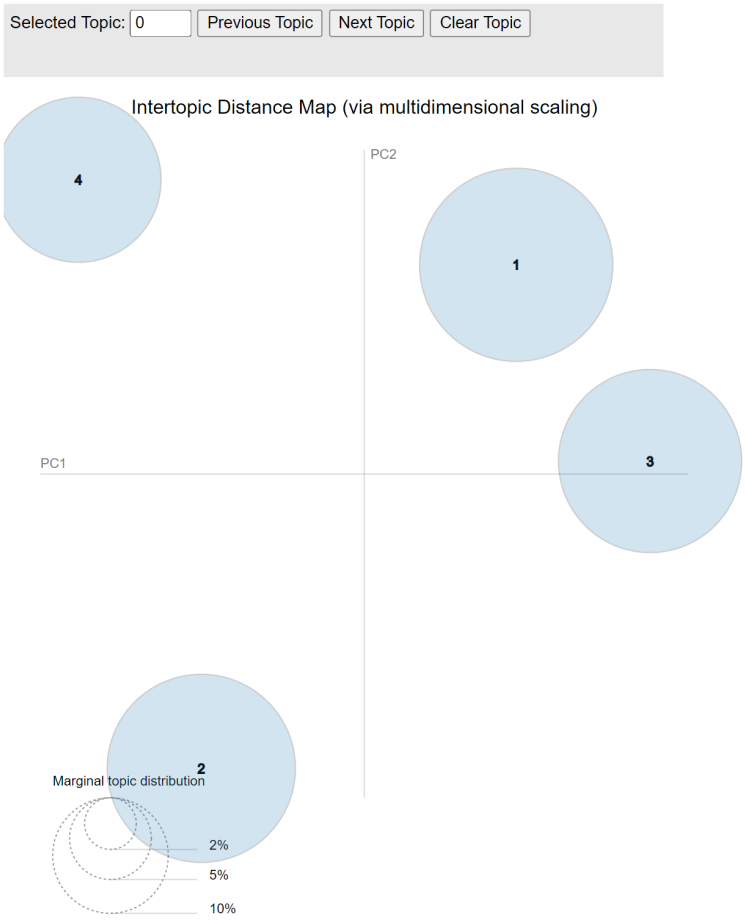


**Figure 3.** Conditional distribution bubble figure
**图 3.** 条件分布气泡图

传转向产品本身解决需求的评价，并再具体到产品的某一个具体的特征的评价。从中可以发现，消费者对网红产品评论的主题主观情感色彩不断减少，越来越趋于理性。因此，第二主成分可以命名为消费理性维度。上图还可分析发现，功能性维度上，评论主题的分布比较平均，左右各两个主题，但是消费理性维度上，三个主题都偏向理性较低的象限。由此可得到结论：网红产品的消费者购买时的消费理性较低，比较容易因为情绪、宣传、包装等主观感受的影响消费。

## 参考文献

[1] 李慧, 胡云凤. 基于动态情感主题模型的在线评论分析[J]. 数据分析与知识发现, 2017, 1(9): 74-82.

[2] 马超. 基于主题模型的社交网络用户画像分析方法[D]: [硕士学位论文]. 合肥: 中国科学技术大学, 2017.

[3] 郭光明. 基于社交大数据的用户信用画像方法研究[D]: [博士学位论文]. 合肥: 中国科学技术大学, 2017.

[4] Robertson, S. (2004) Understanding Inverse Document Frequency: On Theoretica1 Arguments for IDF. *Journa1 of Documentation*, **60**, 503-520. https://doi.org/10.1108/00220410410560582

[5] Breitinger, C., Gipp, B. and Langer, S. (2015) Research-Paper Recommender Systems: A Literature Survey. *International Journal on Digital Libraries*, **17**, 305-338. https://doi.org/10.1007/s00799-015-0156-0

[6] Mikolov, T., Sutskever, I., Chen, K., *et al.* (2013) Distributed Representations of Words and Phrases and Their Compositionality. *Advances in Neural Information Processing Systems* (*NIPS*), 26 p.

[7] Lafferty, J., McCallum, A. and Pereira, F.C.N. (2001) Conditional Random Fields: Probabilistic Models for Segment-

ing and Labeling Sequence Data.

[8] Blei, D.M., Ng, A.Y. and Jordan, M.I. (2003) Latent Dirichlet Allocation. *Journal of Machine Learning Research*, **3**, 993-1022.

[9] Yao, T.F., Cheng, X.W., Xu, F.Y., *et al.* (2008) A Survey of Opinion Mining for Texts. *Journal of Chinese Information Processing*, **22**, 71-80.

[10] Huang, Y.R., Wang, R., Huang, B., Wei, B., Zheng, S.L. and Chen, M. (2021) Sentiment Classification of Crowdsourcing Participants' Reviews Text Based on LDA Topic Model. *IEEE Access*, **9**, 108131-108143. https://doi.org/10.1109/ACCESS.2021.3101565

[11] 吴梦蝶, 唐雁. 基于主题模型的有向社交网络链接预测方法[J]. 西南大学学报(自然科学版), 2014, 36(1): 152-158.

[12] Zhang, N., Liu, R., Zhang, X.-Y. and Pang, Z.-L. (2021) The Impact of Consumer Perceived Value on Repeat Purchase Intention Based on Online Reviews: By the Method of Text Mining. *Data Science and Management*, **3**, 22-32. https://doi.org/10.1016/j.dsm.2021.09.001

[13] Yuan, F.X., Li, M., Liu, R., Zhai, W. and Qi, B. (2021) Social Media for Enhanced Understanding of Disaster Resilience during Hurricane Florence. *International Journal of Information Management*, **57**, Article ID: 102289. https://doi.org/10.1016/j.ijinfomgt.2020.102289