Operations Research and Fuzziology 运筹与模糊学, 2024, 14(1), 279-292
Published Online February 2024 in Hans. https://www.hanspub.org/journal/orf
https://doi.org/10.12677/orf.2024.141027

Hans汉斯

基于文本挖掘的大疆无人机评论情感实证研究

尚雨浩，陈涵怡，金婷婷

南京信息工程大学数学与统计学院，江苏 南京

收稿日期：2023年12月1日；录用日期：2023年12月20日；发布日期：2024年2月22日

摘　要

充分利用好电商平台的文本评论语料，可以挖掘出商品和企业背后的优势和潜在价值。本文通过网络爬虫工具获取大疆无人机在京东商城的网购评论，在对评论语料进行文字预处理的基础上，通过传统的情感词典方法以及机器学习中的伯努利朴素贝叶斯、KNN、SVM对评论语料所含的情感倾向分类并评价。对大疆无人机商品建立LDA主题模型，计算余弦值距离确定最优主题数后更深一步挖掘评论的主题及关注点。发现消费者对于大疆无人机的质量、飞行操纵性、品牌效应、视频拍摄效果、物流配送和配套设施较为关注。最后依据文本挖掘的结果，分析大疆产品的优势，并为生厂商和客户分别提供相关建议。

关键词

文本挖掘，情感倾向分析，机器学习，LDA主题模型

An Empirical Study on Emotional Analysis of Dajiang UAV Comments Based on Text Mining

Yuhao Shang, Hanyi Chen, Tingting Jin

School of Mathematics and Sstatistics, Nanjing University of Information Science & Technology, Nanjing Jiangsu

Received: Dec. 1st, 2023; accepted: Dec. 20th, 2023; published: Feb. 22nd, 2024

Abstract

Making full use of the text comment corpus of e-commerce platform can explore the advantages and potential value behind commodities and enterprises. This paper obtains the online shopping comments of Dajiang UAV in Jingdong Mall through the web crawler tool. Based on the word pre-processing of the comment corpus, this paper classifies and evaluates the emotional orientation contained in the comment corpus through the traditional emotional dictionary method and Bernoulli Naive Bayes, KNN and SVM in machine learning. Establishing LDA theme model for Dajiang

文章引用: 尚雨浩, 陈涵怡, 金婷婷. 基于文本挖掘的大疆无人机评论情感实证研究[J]. 运筹与模糊学, 2024, 14(1): 279-292. DOI: 10.12677/orf.2024.141027

UAV products, calculateing the cosine distance to determine the optimal number of themes, and further explore the themes and concerns of comments. It is found that consumers pay more attention to the quality, flight maneuverability, brand effect, video shooting effect, express delivery and supporting facilities of Dajiang UAV. Finally, according to the results of text mining, this paper analyzes the advantages of Dajiang products, and provides relevant suggestions for manufacturers and customers respectively.

Keywords

Text Mining, Emotional Orientation Analysis, Machine Learning, LDA Topic Model

1. 引言

得益于移动互联网技术的蓬勃发展，移动支付技术的快速普及以及物流配送服务的有效覆盖，网络购物逐渐从一个新颖的概念融入普通人的日常生活。随着网络购物兴起的还有各大电商设置的网络评论系统。对于这些海量的商品评论，它们包含了商品一切有关信息，大到产品质量，功能和特点，小到物流配送，客服态度等等方面。这些评论不仅是消费者对产品最真实的情感反应，也暗含着巨大的信息价值[1]。

无人机作为当下发展前景最火的智能设备之一，不仅广泛应用于目前的军事战争中，其在航拍、农业、植保、快递、测绘等民用领域的功能也愈发重要。但作为耐用型产品，其自身也具有价格昂贵、更新速度快的特点。因此，消费者在购买无人机设备时往往要借鉴他人购物的评论来了解产品的特性及优缺点，以此避免因信息不对等所带来的决策错误。整理、归纳海量的评论信息如果仅用人工阅读难免会遗漏部分信息，对于阅读者而言也过于繁琐，所以对消费者在电商平台留下的评论文本进行挖掘、情感和语义分析具有重要意义[2]。

在对文本挖掘的应用研究可以归纳出，早期研究的对象侧重于专业影评、新闻评述等规范化文本[3]。近些年研究的热点逐渐转向开放度高、逻辑性弱的评论语料，比如旅游服务[4]、手机[5]、电脑[6]等。本文深入挖掘大疆无人机产品的在线评论，丰富和推进了文本挖掘在应用领域的研究现状[7]。

2. 研究方法

2.1. 支持向量机

支持向量机(Support Vector Machines, SVM)，是一种广义线性分类器，它基于监督学习将数据分类为二进制模式。在进行分类决策时所划分的类别界限是根据训练样本不同类别之间距离最远的一个超平面，从而将分类问题变成了传统的二次线性规划。当结构风险最小化时，经验风险和置信区间可以同时最小化。简单来说就是当两类样本线性分离时，在原始空间中找到最优的分类超平面。以下是一些具体的想法：

如图 1 所示，当训练数据线性可分时，将训练数据 $T=\left\{\left(x_1,y_1\right),\left(x_2,y_2\right),\cdots,\left(x_N,y_N\right)\right\}$，其中 $x_i\in R^n$，$y_i\in\left\{-1,1\right\}$，构造并求解最优化问题：

$$\min\frac{1}{2}\|\omega\|^2 \tag{1}$$

$$s.t.\ y_i\left(\omega\cdot x_i+b\right)-1\geq 0 \tag{2}$$

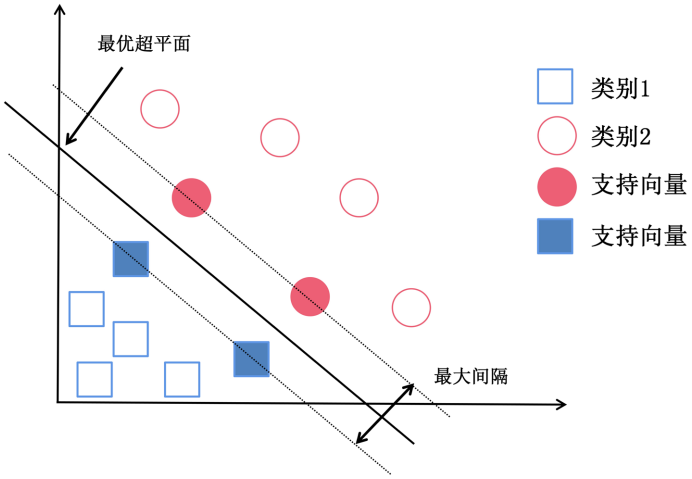

Figure 1. Schematic diagram of linearly separable support vector machine classification
图 1. 线性可分时支持向量机分类示意图

如果遇到训练数据线性不可分的情况时，我们可以引入松弛变量，通过非线性映射把原始空间映射到一个高纬度的特征空间。在特征空间内找到一个最理想的超平面，实现分类。但在现实处理过程中，由于噪音的存在，我们需要借助核函数来实现这一目的。常见的核函数有如下几种：

$$\text{多项式核函数：} k\left(x_i, x_j\right) = \left(x_i^{\mathrm{T}} x_j\right)^d \tag{3}$$

$$\text{高斯核函数：} k\left(x_i, x_j\right) = \exp\left(-\frac{\left\|x_i - x_j\right\|^2}{2\sigma^2}\right) \tag{4}$$

$$\text{Sigmoid核函数：} k\left(x_i, x_j\right) = \tanh\left(\beta x_i^{\mathrm{T}} x_j + \theta\right) \tag{5}$$

2.2. K 近邻算法

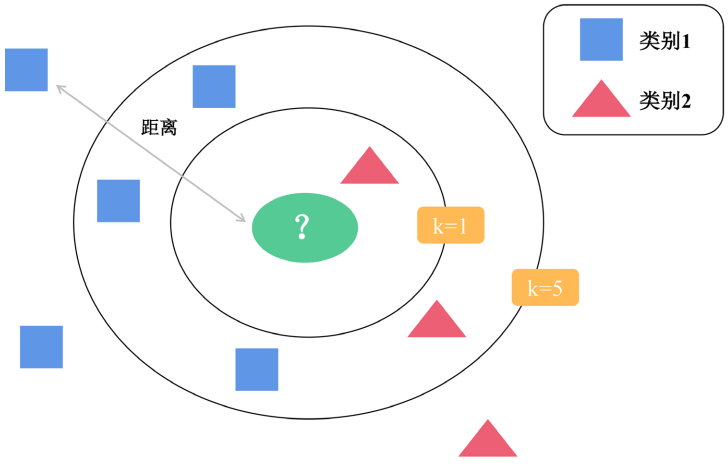

Figure 2. KNN algorithm classification diagram
图 2. KNN 算法分类示意图

K 近邻(K-Nearest Neighbor, KNN)，是□种经典的机器学习算法，常用于分类和聚类。该方法的基

本思路可以用图 2 说明：分析绿色椭圆形代分类样本所在的特征空间情况，当 k 选定 1 时，距离其最近的一个样本类别是红色三角形，属于类别 2，则此时待分类的绿色椭圆形也属于类别 2。如果选定 k 取 5，距离其最近的五个样本为三个属于类别 1 的蓝色方块和 2 个属于类别 2 的红色三角形，则待分类样本此时属于类别 1。由于 KNN 算法在分类时需要明确待分类样本周边其他样本的类别所属情况以便后续分析，因此 KNN 属于有监督的算法。

KNN 算法的具体步骤如下：

① 将没有分类的文本用它的特征向量表示；

② 计算没有分类文本和已知文本的距离，通常用 Sim_hash 算法来计算距离，定义如下：

$$Sim\left(d_i,d_j\right)=\frac{\sum_{k=1}^{M}w_{ik}\cdot w_{jk}}{\sqrt{\sum_{K=1}^{M}w_{ik}^2}\sqrt{\sum_{K=1}^{M}w_{jk}^2}} \tag{6}$$

其中 d_i 为待分类文本的特征向量，d_j 表示 j 类的中心向量，M 是向量的维数。

③ 找到最为接近的 K 个样本，并获取它们的标签；

④ 在最接近的 K 个近邻样本中，对不同类别样本赋予不同的权重并进行计算，计算公式如下：

$$p\left(X,C_j\right)=\begin{cases}1 & \text{如果}\sum_{d_i\in KNN}Sim\left(x,d_i\right)y\left(d_i,C_j\right)-b\geq 0\\ 0 & \text{其他}\end{cases} \tag{7}$$

其中 $Sim\left(x,d_i\right)$ 是距离计算公式，b 为距离阀值。y 取值为 0 或 1，如果 $d_i\in C_j$，则 $p\left(X,C_j\right)=1$，反之则为 0。

2.3. 伯努利朴素贝叶斯

伯努利朴素贝叶斯 BNB (Bernoulli Naïve Bayes)是一种专门用于处理二项分布的朴素贝叶斯模型。朴素贝叶斯基于贝叶斯定理与特征条件独立假设的分类方法，其中朴素指的就是条件独立。朴素贝叶斯在分类的时候不是直接返回分类，而是返回属于某个分类的概率。

朴素贝叶斯法对条件概率分布作了条件独立性的假设：

$$P\left(X^{(1)}=x^{(1)},\cdots,X^{(n)}=x^{(n)}\middle|y=c_k\right)=\prod_{j=1}^{n}P\left(X^{(j)}=x^{(j)}\middle|y=c_k\right) \tag{8}$$

在利用朴素贝叶斯法进行分类时，对给定的输入 x，可以计算后验概率分布 $P\left(c_k\middle|x\right)$：

$$P\left(c_k\middle|x\right)=\frac{P\left(x\middle|c_k\right)P\left(c_k\right)}{\sum P\left(x\middle|c_k\right)P\left(c_k\right)}=\frac{\prod_{j=1}^{n}P\left(X^{(j)}=x^{(j)}\middle|y=c_k\right)P\left(c_k\right)}{\sum\prod_{j=1}^{n}p\left(X^{(j)}=x^{(j)}\middle|y=c_k\right)P\left(c_k\right)} \tag{9}$$

其中上式中对所有 c_k 分母都相同，因此可以化为：

$$h^*\left(x\right)=\arg\max_{c_k\in y}\prod_{j=1}^{n}P\left(X^{(j)}=x^{(j)}\middle|y=c_k\right)P\left(c_k\right) \tag{10}$$

根据伯努利分布的表达式 $P\left(X=x\right)=px+\left(1-p\right)\left(1-x\right)$，伯努利贝叶斯模型的目标函数表达式为：

$$P\left(X=x\right)=px+\left(1-p\right)\left(1-x\right)y=\arg\max_{c_k\in y}\prod_{j=1}^{n}\left(x^jP\left(X^{(j)}=x^{(j)}\middle|y=c_k\right)+\left(1-x^j\right)\left(1-P\left(X^{(j)}=x^{(j)}\middle|y=c_k\right)\right)\right) \tag{11}$$

2.4. LDA 主题模型

LDA (Latent Dirichlet Allocation)主题模型[8]表示潜在狄利克雷分配模型，是一种潜在语义挖掘模型。在模型的假设条件下，文档、主题和分词均服从多项分布。该模型对每个文档进行主题提取，然后按照主题的分布进行分词，对每个文档、主题和分词都进行重复式的主题提取和分词。LDA 模型采用的是词汇布袋模式，因此没有考虑单词间出现的先后顺序，只对单词是否出现进行研究。以此，利用每个文档所抽取的主题和分词数量，每篇文档都可以降维并通过一个词频向量表示出来，实现了文本到向量的转化。LDA 模型结构如图 3 所示：

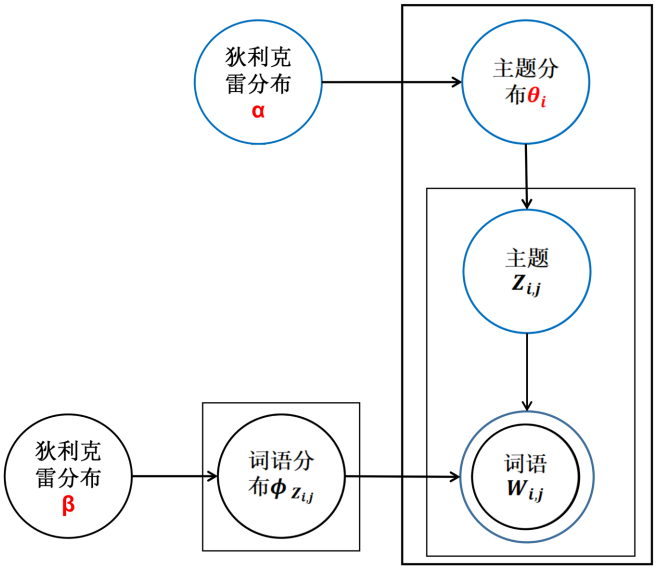

Figure 3. LDA model structure diagram
图 3. LDA 模型结构图

LDA 模型结构如上图所示。α 和 β 是文档主题和主题词语的 Dirichlet 函数的先验参数；θ 是主题在文档中的多项分布的函数，其服从超参数为 α 的 Dirichlet 先验分布；ϕ 是词语在主题中的多项分布的参数，其服从超参数 α 的 Dirichlet 先验分布。LDA 模型假设每条评论是所有的主题按照一定比例组合而成的，而组合比例服从于多项分布，记为：

$$Z|\theta \sim Multi(\theta) \tag{12}$$

每个主题词典是由所有的分词按照比例组合而成，组合比例也遵循多项分布，记为：

$$W|Z,\phi \sim Multi(\phi) \tag{13}$$

在评论 d_j 条件下生成词 w_i 的概率表示为：

$$P(d_j|w_i) = \sum_{s=1}^{K} P(w_i|z=s) \cdot P(z=s|d_j) \tag{14}$$

其中，$P(w_i|z=s)$ 表示词 w_i 属于第 s 个主题的概率；$P(z=s|d_j)$ 表示第 s 个主题在评论 d_j 中的概率。

3. 数据来源及预处理

3.1. 数据的选取

本文选取的是时下热门的科技产品——民用无人机的电商评论。在该领域国产品牌大疆可谓遥遥领先

市场同级别竞争对手。为了探究消费者心中对大疆产品的关注点，以及大疆自家产品的黑科技，选取大疆目前在售的三款主流产品，分别是大疆 Mini，Air 和 Mavic。价格涵盖了从基础款 3000 元到顶配版 15,000 元。大疆品牌的无人机因为其高超的性能和优良的做工，价位一般略高于普通厂商同级别的无人机产品。得益于此，购买大疆产品的消费者往往有着更加专业的无人机使用背景，对其的需求度较高。优质的买家往往才会产生优质的评论。过于便宜的无人机更多被用作小朋友的玩具，不具备优质性能，存在的产品缺陷过高。而更为专业的无人机产品受众面太窄，无法从大众视角下的评论探究其科技含量。

评论爬取的内容主要包括：用户名，打分星级，评论内容，商品参数，评论时间。通过设置爬取规则，采集了大疆 Mini、Air、Mavic 三种商品共 17,149 条在线评论。综合评价星级和具体内容，对文本语料进行打标签，其中正面情感倾向评论 15,171 条，负面评论 1978 条。部分商品评论数据如表 1 所示：

Table 1. Partial product review data table
表 1. 部分商品评论数据表

用户 ID	评价星级	评价内容	商品参数	评论时间
y***狸	star5	大疆无人机一无继往的牛，做工非常好，设计合理，折叠起来体积很小放到包里一点也不占地方而且螺旋桨也不用卸下来一折刚好和机体完美贴合不会刮蹭产品质量稳定飞行很顺滑坚决好评。	大疆 DJI 御 Mavic 3 Pro	2023/11/13 19:42:19
浅***g	star5	东西已经收到！手感不错！用料扎实！质量很好！平常用基本没有问题！非常值得信赖的品牌，质量有保障，价格还亲民，相当满意快递很快，质量棒极了，建议购买外观干净漂亮，做工很不错，没发现什么瑕疵，个人感觉大小也挺合适，找了很久选的这款，大爱！	大疆 DJI 御 Mavic 3 Pro	2023/11/13 19:54:34
****a	star4	送货较快，毕竟是高精密电子产品，但是，包装没有必要保护，导致运输途中无人机外包装盒有压损，不太满意。	大疆 DJI Air 3 套装	2023/11/14 20:39:33

3.2. 数据的清洗

清洗的工作不仅包括删除采集时重复的内容，还包括语句的精简压缩。例如像“好好好！”、“不错不错不错”这样的短评可以压缩成“好！”、“不错”。针对系统的默认评论“此用户未做出评论”也需要删除。此外，部分评论还带有非规范的表意不明的符号，例如“^_^”、“o(∩_∩)o”等也需要剔除。文本数据清洗完成后，共有 11,751 条有效数据。

3.3. 分词和去停用词

本文选用的是 python 中的 gensim 模块的 jieba 分词包进行这一操作。jieba 分词的本质是一个概率语言的分词模型，本文将大疆无人机的 11,751 条评论读入 jieba 自带的 txt 词典后，共被分为了 418,891 个词语。

分词过后，原本的评论语句会被拆分成单个独立的单词，句子依旧保持完整性。可被拆分的句子中留有许多不作语句成分的词汇，包括副词、介词、连词等。它们在句子中没有表示具体的含义，仅仅是为了语句的通畅，例如“的、了、呢、然而、但是”等词。为了避免后续对文本语料处理时残留过多的噪音，我们需要对这些没有实际意义的词进行剔除。本文借助的是哈工大停用词，在对无人机商品评论进行剔除停用词和标点等噪音后，得到不含停用词的分词表，共 182,022 条。

4. 无人机评论文本挖掘

4.1. 情感词典分类效果探究

使用情感词典对文本语料进行情感倾向分析是一种直截了当且质朴的分析方法[9]。它的基本做法是根据情感词典事先编辑好的词条对待分析文本情感词进行匹配并赋值。积极词条赋予正值，消极赋予负值。然后根据对所有文本赋值的结果进行汇总并计算最后的得分。最后依据事先划定好的阀值，对所有打分的文本分类。本文采用表 2 所示的三款基础词典构建情感词典。

Table 2. Emotional dictionary construction table
表 2. 情感词典构建表

词典名称	积极词汇个数	消极词汇个数
清华大学李军中文褒贬义词典	5567	4469
NTUSD 简体中文情感词典	2810	8274
知网 Hownet 词典	4566	4370

对三款词典汇总，去重等处理后得到的本文所使用的基础词典共含有 23,633 个情感次，其中积极词条 10,219，消极词条 13,414 个。

构建好基础情感词典后，就可以遍历所有语句然后对情感词进行赋分。本文定义 weight，对正面情感词赋值为+1，负面情感词赋值为−1。此外，对于情感词的评判还需要设置 amend_weight 来修正情感倾向。因为在汉语表达方式中，双重否定表肯定是一种常见的表达方式，如有多重否定，那么奇数否定是否定，偶数否定是肯定。所以我们处理时要看该情感词前 2 个词，来判罚否定的语气。如果在句首，则没有否词，如果在句子的第二次词，则看前 1 个词，来判断否定的语气。部分否定词处理结果如下表 3 所示：

Table 3. Partial assignment table of negative words
表 3. 部分否定词赋分表

index_content	word	nature	content_type	weight	amend_weight
1	简单	a	pos	1.0	1.0
2	不错	a	pos	1.0	1.0
3	值得	v	pos	1.0	1.0
4	慢	a	neg	−1.0	−1.0

计算完每个词条的情感得分后，就可以对词条所在语句计算其总体得分。我们设定 0 为情感阀值，将语句情感得分大于 0 的归为正面评论，小于 0 的归为负面评论，等于 0 的定义为中性评论。最终得到 10,172 条情感得分不为 0 的评论语句。部分语句得分情况如表 4 所示：

Table 4. Score table of some comment sentences
表 4. 部分评论语句得分表

index_content	amend_weight	ml_type
1	19.0	pos
2	5.0	pos
3	1.0	pos
4	7.0	pos
5	−3.0	neg

为了评估情感词典对文本语句的分类效果，引出文档分类的基本定义：

1) Truepositives (TP)：正样本分类后被预测为正样本的个数。

2) Falsepositives (FP)：正样本分类后被预测为负样本的个数，在假设检验中也常常被称为第一类错误。

3) Truenegatives (TN)：负样本分类后被预测为负样本的个数。

4) Falsenegatives (FN)：负样本分类后被预测为正样本的个数，在假设检验中也常常被称为第二类错误。

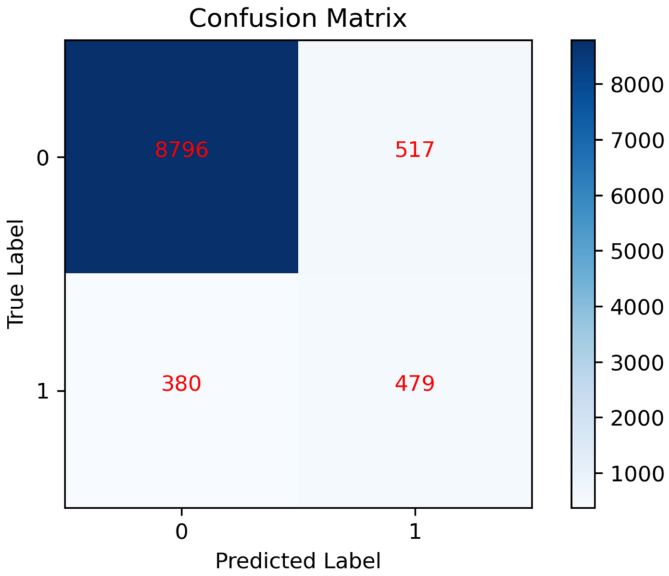

Figure 4. Confusion matrix diagram of the prediction results of the emotional dictionary method
图 4. 情感词典法预测结果的混淆矩阵图

情感词典法预测结果的混淆矩阵如图 4 所示。由混淆矩阵可知，构建词典法对评论文本分类的正确率为 91.18%。由此可见，利用情感词典法对评论文本情感分类的效果还是比较理想的。如果直接利用情感词典法对每条评论进行分类并标注，就可以省去人工手动标注的繁琐。

4.2. 机器学习分类效果比较

4.2.1. 实验步骤

数据来源：在通过构建词典法获得分类标签的文本数据后，把所有评论语料以七三比划分成训练集和测试集，接着用监督学习对其分类。

数据处理：在对文本语料用 jieba 分词和去除停用词后，将分词结果导入训练好的 word2vec 模型中提取特征。

实验与评估：利用伯努利朴素贝叶斯、KNN、SVM 三种监督学习方法，训练过后对测试集文本分类，根据给定的评估指标对分类器效果进行评价。

4.2.2. 分类器性能评估

模型建立好以后，我们要根据分类的结果建立一套尽可能客观的模型评价系统，其目的是评判分类器的分类准确程度。最重要也最常见的评估指标为准确率(Accuracy)，其表示如下：

$$\text{Accuracy}=\frac{\text{TP}+\text{TN}}{\text{TP}+\text{FP}+\text{TN}+\text{FN}} \tag{15}$$

在 python 中可以通过 scikit-learn 包中的 accury_score 来计算模型分类的准确率。以本文的二分类为例，如果样本数据正、负不均衡，我们用查准率(PRE)、召回率(REC)、F1 值这三个指标来衡量分类器效果的优劣会更为妥当。引出本文所用的三项评价指标的定义：

查准率 PRE 主要用来预测正面结果，真正的正面样本在所有预测为正面的样本中所占比例：

$$\text{PRE}=\frac{\text{TP}}{\text{TP}+\text{FP}} \tag{16}$$

召回率 *REC* 针对样本，表示所有预测为正面的样本中有多少是真正预测正确的：

$$\text{REC}=\frac{\text{TP}}{\text{TP}+\text{FN}} \tag{17}$$

F_1 是 Accuracy 和 REC 的调和平均数，是 F_{beta} 指标中 *beta* 为 1 时的值：

$$F_{beta}=\left(\beta^2+1\right)\frac{\text{Accuracy*REC}}{\beta^2\left(\text{Accuracy}+\text{REC}\right)}=\frac{\left(\beta^2+1\right)TP}{\beta^2\left(FN+FP\right)+\left(1+\beta^2\right)TP} \tag{18}$$

Beta 为 1 时， F_1 定义如下：

$$F_1=\frac{2\cdot\text{PRE}\cdot\text{REC}}{\text{PRE}+\text{REC}} \tag{19}$$

本文分别取 1000、2000 和 3000 条评论语料分别进行三次分类测试，按照 7:3 划分训练集和测试集。并计算分类器在测试集上评估 100 次评价指标的均值，分类器的三种评价值如表 5 所示：

Table 5. Classifier result evaluation table
表 5. 分类器结果评估表

语料	BNB			KNN			SVM		
数目	PRE	REC	F1 值	PRE	REC	F1 值	PRE	REC	F1 值
1000	87.33%	90.45%	88.86%	84.33%	89.85%	87%	89.83%	95.64%	92.64%
2000	89.66%	91.14%	90.39%	86.89%	93.82%	90.22%	91.11%	95.73%	93.36%
3000	90.66%	94.33%	92.46%	87.16%	93.74%	90.33%	91.66%	97.48%	94.48%

从表 5 可以看到，随着语料数目的增多，三种分类器的分类效果均在逐渐变好。这是因为训练集内部的用于训练的特征语句会随着全部语料数目的增加而增加，模型更完善，分类器训练的精准度就会得到提高。当在语料实验投入的文本最多时，SVM 分类器的 REC 和 F_1 值均最高。这说明随着语料数目增多，SVM 分类器在对语料分类处理时犯第二类错误的概率最小，且总体分类效果最好。这与朱婧[10]在对新能源汽车评论文本情感分类测试得到的结论一致。

4.3. 词云可视化

对大疆无人机商品评论进行词云可视化分析，积极、消极词云图分别如图 5 和图 6 所示。从正面评论的名词词云图我们可以看出，消费者在无人机的灵敏度、稳定性、操作、外观、难易等词词上持正面情感态度。据此我们可知，大疆无人机的黑科技集中体现在飞机的飞行性能上。其灵敏度较高、飞行性能以及拍摄视频的稳定性极高，拥有出色的做工和外观。此外，大疆无人机操作起来也相对容易，对于

新手十分友好。

Figure 5. Word cloud of nouns for positive comments
图 5. 积极评论的名词词云图

Figure 6. Word cloud of nouns for negative comments
图 6. 消极评论的名词词云图

对于大疆无人机的消极评论，我们可以发现这些评论的集中点似乎并没有集中在产品本身，而是集中在相关配套设施和购买体验上。甚至不错这个形容词在负面评价的出现频率也较高，说明即使消费者在商品评论中给出了消极的结果，但对于产品的总体评价却趋于认同。此外，我们也能看到消费者吐槽点主要在电池、客服、飞行信号、以及视频拍摄反馈的眼镜上。

4.4. LDA 主题挖掘

4.4.1. LDA 模型最优主题个数的计算

确定 LDA 主题数目有多种方法，本文通过计算不同主题间的余弦值来寻找和确定最优主题个数[11]。计算余弦值可以用较少的计算就可以确定主题的最优个数，而且不需要人工多次调试参数。

余弦相似度描述的是 n 维空间中两个维数为 n 的向量之间夹角的余弦。它是两个向量的点积除以两个向量的长度(或幅度)的乘积。定义 A 是 $(A_1, A_2, \cdots, A_n)$，B 是 $(B_1, B_2, \cdots, B_n)$。二者之间的余弦值距离计算如下：

$$\text{similarity}(A,B) = \frac{A \cdot B}{\|A\| \times \|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \times \sqrt{\sum_{i=1}^{n} B_i^2}} \tag{20}$$

由余弦值的定义我们可知，余弦相似度的值越小，则代表它们相似程度越高，所对应的模型最好。具体流程如下：

① 先选定一个数值 k 作为主题的个数，计算这种情况下 similarity(A,B) 大小。

② 改变 k 值的大小，重新训练，再次计算 similarity(A,B) 大小。

③ 重复②，直到找出使 similarity(A,B) 降至最低时的 k 值。

图 7 是正面评论和负面评论余弦值曲线图：

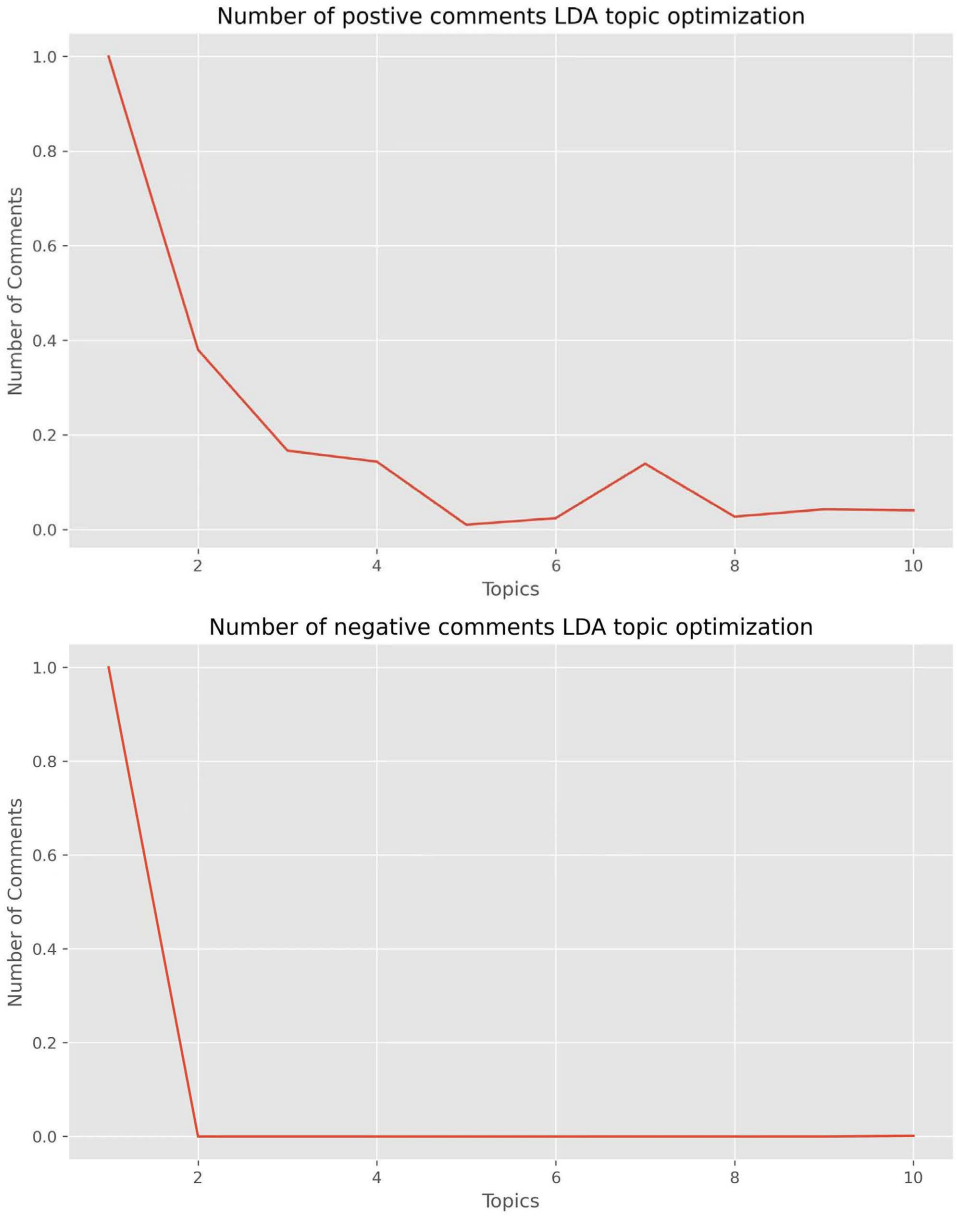

Figure 7. LDA topic number optimization diagram
图 7. LDA 主题数寻优图

由图可知，主题间平均余弦值降至最低时，正面评论语料的主题数 k 取 5，负面评论语料的主题数 k 取 2~9。因此，对正面评论数据做 LDA，可以选择主题数为 5，负面主题数可选为 2。

4.4.2. 大疆无人机好评集和差评集的主题分析

Table 6. LDA topic model analysis results of DJI UAV praise collection
表 6. 大疆无人机好评集 LDA 主题模型分析结果表

Topic1	Topic2	Topic3	Topic4	Topic5
做工	不错	感觉	视频	购买
质量	操作	好评	画质	体验
高	稳定性	品牌	试飞	收到
外观	飞	产品	拍摄	很快
外形	灵敏度	特色	清晰	高度
京东	喜欢	满意	好	小巧
难易	值得	国产	携带	价格
不错	稳定	性价比	效果	综合
飞行	手	东西	值得	拍
电池	新手	棒	体验	飞行

大疆无人机好评集和差评集的 LDA 主题模型分析结果如表 6 和表 7 所示。根据表 6 大疆无人机好评集 LDA 主题模型分析结果来看，主题一“质量”、“外形”、“做工”、“不错”等词反映出消费者对于大疆品牌无人机的整体评价是满意的，反应大疆产品质量高，外形优美，做工用料足，博得了消费者一致认可。主题二中关键词“操作”、“稳定性”、“灵敏度”、“飞行”则反映消费者对于大疆无人机的飞行体验，操纵手感和日常使用上十分认可，说明大疆无人机除了硬件用料足，在优化客户日常使用体验，对飞行器飞行的软件优化做的也很到位。主题三中“品牌”、“产品”、“特色”、“国产”等一干词汇侧面表现出购买者看中大疆的品牌效应，追求国产品牌。主题四“视频”、“拍摄”、“清晰”等表现出消费者对于大疆无人机拍摄质量的赞扬。主题五的词汇比较综合，体现出消费者对于大疆产品综合体验上的感觉还是不错的。

Table 7. LDA topic model analysis results of DJI drone negative reviews collection
表 7. 大疆无人机差评集 LDA 主题模型分析结果表

Topic1	差	包装	新手	不好	售后	感觉	飞行	时间	快递	原因
Topic2	不想	价格	被迫	客服	玩	第一次	旅程	电池	京东	返航

根据表 7 大疆无人机差评集 LDA 主题模型分析结果来看，主题一集中反映的是双 11 节日原因对于大疆无人机出售、供货、配送的影响。结合爬取的评论日期我们结合有关背景可以了解到，双 11 是物流高峰期，部分店家商品出现预售无货、物流慢，甚至快递包装破损等情况，故而引起消费者负面情绪。主题二中“价格”、“电池”、“京东”则集中反应大疆在对优化用户购买体验上做的不够。产品性能虽好，但是在产品赠品、平台出售等细节上需要继续提高。

5. 结论与建议

本文得出的结论主要有：

利用构建情感词典的方法对语句情感词进行打分并判别情感倾向准确率较高，在面对海量没有标签的文本时，可以构建情感词典计算语句情感值。这种方式有效避免了传统人工标注的麻烦，在面对海量

文本数据时较为适用。

本文选取大疆主流无人机产品在线评论数据17,149条，清理过后有效文本数量为11,751条。在比较三种机器学习方法对文本情感分类时，SVM分类效果最优，准确率最高，REC和F1值显著高于KNN和伯努利朴素贝叶斯算法。因此，在使用SVM对数据分类时可以提高对负面评论分类的准确率，从而有效避免犯第二类错误的概率。

从大疆无人机的LDA主题模型分析结果来看，大疆品牌在消费者心中已经留下了：质量够硬、外形炫酷、操纵体验感极佳、飞行和视频拍摄质量稳定等良好印象。可见，大疆品牌能引领全球民用无人机发展的浪头绝非靠营销和宣传，其口碑和品牌形象已经深深得到了消费者的认可[12]。

5.1. 对大疆企业的建议

优化产品零部件以及消费者购买时的消费体验。差评反映较多的问题集中在无人机的电池续航、信号强度还有附赠的智能眼镜上。对于这一类产品虽然不是大疆厂商亲自生产的商品，但仍要保证质量，精益求精。商家的优良形象体现在方方面面，切不可在最后的环节出现差错。

管控生产线，保证产品出货以及供应。今年的双11热度虽较往年有所下降，但双11购物节期间的库存、物流压力仍是本次差评集主要反应的问题。拓展生产渠道，保证热销产品在购物节期间的供应，让顾客少些等待[13]，这些是未来大疆需要提升的方向。做好售后以及退换货服务。本次差评集中出现频率较高的还有“包装”、“京东”、“客服”等字眼。如果说产品反映了品牌的科技形象，而售后更能反映出品牌的管理和负责任形象。大疆若想要成长为巨头公司还需要注重消费者的需求，只有耐心、细致、完善细节才能彻底赢得消费者的支持。

5.2. 对消费者的建议

大疆无人机的产品综合性能值得信赖。从对评论文本的分析可以看到，几乎没有针对大疆无人机本身的负面评价。大疆确实是靠科技水平攻占市场，而非擅长营销和包装的虚伪公司。而积极评论最主要集中在质量、稳定性和灵敏度上面。对于飞行器飞行有一定要求的买家可以尝试入手。

对于电池续航的问题留有争议，需要仁者见仁智者见智。无论是在好评集还是差评集，电池都是它们的常客。可见不同消费者对于相同事物的评价标准不同。如果想要客观评价，最好查询有关电池供应商对于所生产电池的具体参数。

参考文献

[1] 杨立公, 朱俭, 汤世平. 文本情感分析综述[J]. 计算机应用, 2013, 33(6): 1574-1578+1607.

[2] 杨梦琳, 卢益清. 基于在线评论的生鲜电商顾客满意度分析研究[J]. 中国物流与采购, 2022(6): 44-46.

[3] Pang, B., Lee, L. and Vaithyanathan, S. (2002) Thumbs up? Sentiment classification Using Machine Learning Techniques. *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing*, Volume 10. Stroudsburg, 6 July 2002, 79-86.

[4] 蒋亚丽, 李长军. 基于文本数据挖掘影响乘客满意度的因素[J]. 数据挖掘, 2019, 9(3): 88-95.

[5] 张琰, 朱燕翔, 郑桂玲. 基于网购评论文本挖掘的手机类产品属性评价研究[J]. 现代商贸工业, 2018, 39(22): 49-51.

[6] 李清镇. 基于文本挖掘的笔记本电脑网评分析[D]: [硕士学位论文]. 兰州: 兰州财经大学, 2019.

[7] 房文敏, 张宁, 韩雁雁. 在线评论信息挖掘研究综述[J]. 信息资源管理学报, 2016, 6(1): 4-11.

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

[9] 汪韬, 张再跃. 面向电影评论的情感词典构建方法研究[J]. 计算机与数字工程, 2022, 50(4): 843-848.

[10] 朱婧. 基于文本挖掘的新能源汽车在线评论情感分析[D]: [硕士学位论文]. 重庆: 西南大学, 2023.
[11] 陈俊宇. 基于文本挖掘的在线评论应用研究[D]: [硕士学位论文]. 武汉: 湖北工业大学, 2020.
[12] 洪佳滢, 孟子锐. 大疆品牌资产调研及品牌发展策略研究[J]. 国际公关, 2023(11): 137-139.
[13] 胡迪. 基于文本挖掘和情感分析的物流客户满意度测算研究[J]. 物流技术与应用, 2022, 27(3): 158-161.