厚德 求真 励学 笃行

产品评论文本挖掘以及综合应用

成员：王钰翔

甘逸民

赵红玉

➡ 汇报：甘逸民

目录

CONTENT

1 项目背景

XDU

传统分析限制

传统方法难以深度挖掘评论中的多层信息，通常只提供浅层次的分析结果。

综合分析方法

采用情感分析、聚类分析和LDA主题分析相结合的方法，进行服装产品消费者评论综合分析。

综合分析价值

提供全面、深入的市场洞察与决策支持，帮助企业更好地满足消费者需求。

方法启发与借鉴

结果可为其他领域的消费者评论分析提供借鉴和参考。可以进一步拓展分析方法，更准确地洞察市场趋势。

近年来互联网和电子商务蓬勃发展，消费者有在线购买产品的趋势。
爆发增长的消费者评论成为宝贵信息资源，如消费者观点、感受和使用经验等。

2 数据探索

XDU

数据集简介

包含了关于服装产品的全面评论集合，对于多标签分类研究来说是一项宝贵的资源。

Title：评论标题

Review：评论内容

Cons_rating：评价评级

Cloth_class：服饰类型

Materials：布料类型

Construction：布料结构

Color：颜色

Finishing：含义未知，暂且忽略

Durability：耐用性

数据收集来源

本数据集来源于Kaggle平台，作者是来自Telkom University的Nadhif Girawan。

原始数据集共有49338条，9个特征变量。

数据集特征

每个数据条目都被标注了相关标签，让研究人员可以探索服装产品的多个方面。

数据集中的评论提供了丰富的观点和意见，有助于开发稳健的分类模型，准确预测服装项目的多个方面。

2 数据探索

XDU

2.1 读取数据

使用 pandas 读的数据文件，指定了逗号作为分隔符。

打印输出**数据集的前几行**，以便初步了解数据的结构和内容。

Title	Review	Cons_ratir	Cloth_clas	Materials	Constructi	Color	Finishing	Durability
	Absolutely	4	Intimates	0	0	0	1	0
	Love this (	5	Dresses	0	1	0	0	0
Some majo	I had such	3	Dresses	0	0	0	1	0
My favorit	I love, lov	5	Pants	0	0	0	0	0
Flattering :	This shirt i	5	Blouses	0	1	0	0	0

2.3了解更多信息

数据集的**行数和列数和概述信息**

```
df.shape
```

```
(49338, 9)
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 49338 entries, 0 to 49337
Data columns (total 9 columns):
 #   Column        Non-Null Count  Dtype  
---  ------        --------------  -----  
 0   Title         45373 non-null  object 
 1   Review        48509 non-null  object 
 2   Cons_rating   49124 non-null  float64
 3   Cloth_class   49322 non-null  object 
 4   Materials     5741 non-null   float64
 5   Construction  5743 non-null   float64
 6   Color         5742 non-null   float64
 7   Finishing     5737 non-null   float64
 8   Durability    5734 non-null   float64
dtypes: float64(6), object(3)
memory usage: 3.4+ MB
```

3 数据预处理

XDU

3.1 缺失值分析

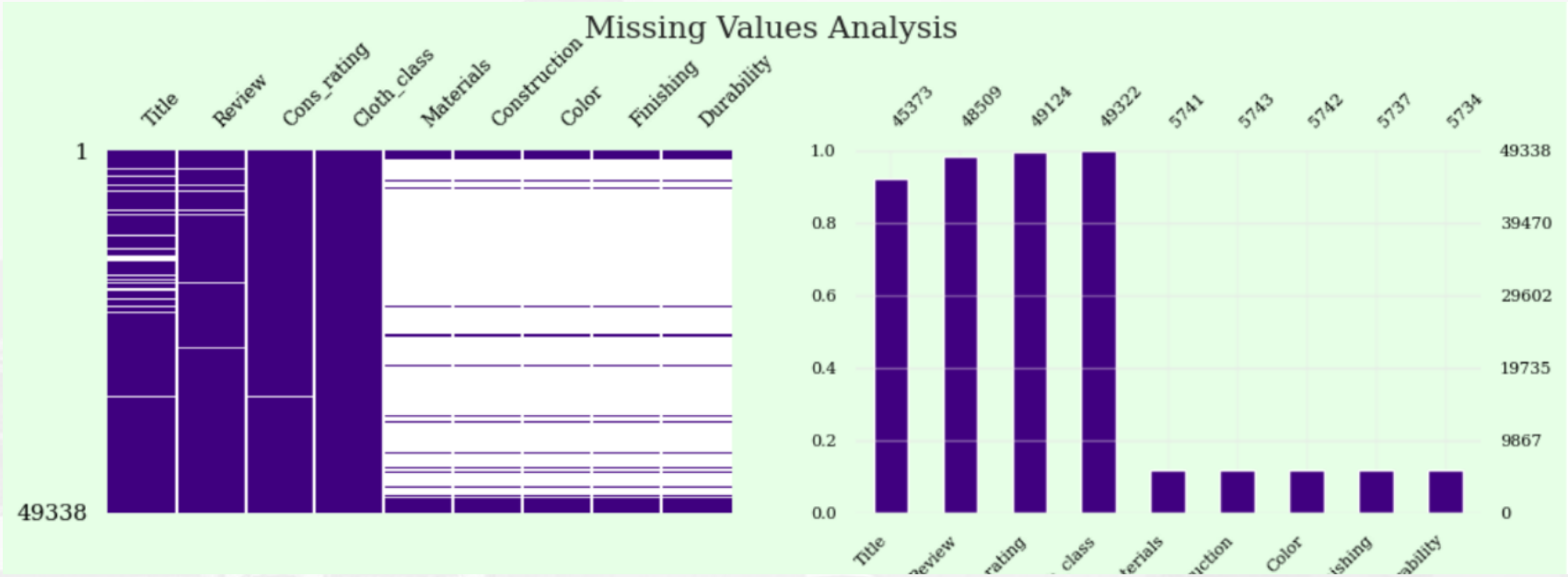
缺失值可视化图表：

矩阵图、柱状图

紫色部分表示有数据

白色部分表示缺失的数据

较多缺失值



3.2 缺失值处理

用数字 0 填充了数据集中的缺失值。

对数据进行了**去重操作**。

```
Data columns (total 9 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   Title         48217 non-null  object
 1   Review        48217 non-null  object
 2   Cons_rating   48217 non-null  float64
 3   Cloth_class   48217 non-null  object
 4   Materials     48217 non-null  float64
 5   Construction  48217 non-null  float64
 6   Color         48217 non-null  float64
 7   Finishing     48217 non-null  float64
 8   Durability    48217 non-null  float64
dtypes: float64(6), object(3)
```

4 数据分析

XDU

4.1 评论词云图

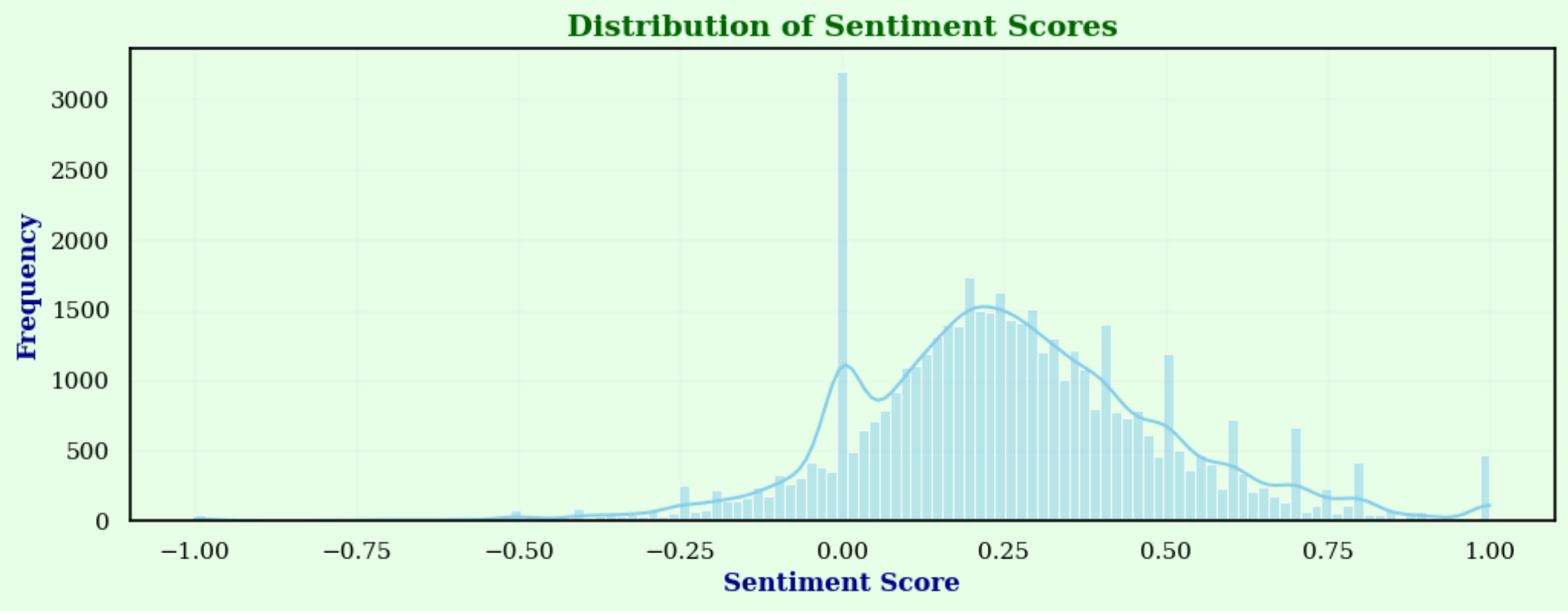
1.消费者最常提及“**dress**”和“**fit**”，表明他们对服装的款式和合身性高度关注。

2. 整体来看，这个词云图反映了消费者在评价服装时，重点关注**款式、合身性、舒适度、面料质量和颜色**等方面。这些信息对商家的产品开发和营销策略具有重要参考价值。

4 数据分析

4.2 情感分析

XDU



1.情感得分分布：评论情感得分从 **-1 到 1**，大量评论集中在 0 附近，表明中性评论占较大比例。**2.正面与负面情感**：正面情感分布集中在 0 到 0.5 之间，显示**大部分评论偏正面**，而负面情感评论相对较少且分散。**3.总结：**整体来看，评论以中性和正面情感为主，负面评论较少，反映出消费者对服装的**总体满意度较高。**

4 数据分析

4.3 数值变量相关系数热力图

1.热力图概述：颜色显示变量间的相关性，蓝色正相关，红色负相关，深浅表示强度。

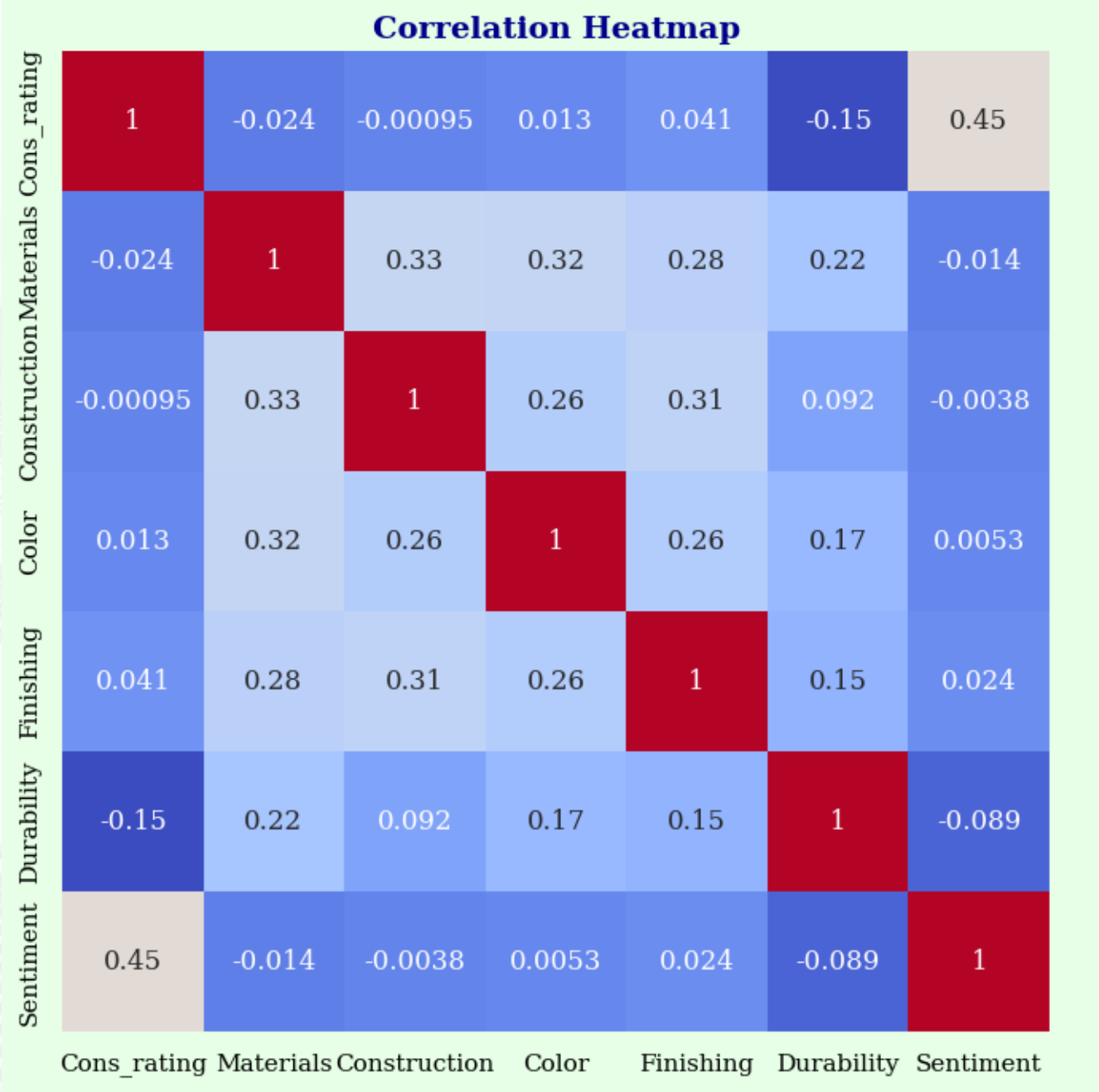
2.关键发现：

情感与耐用性、材料、完成度正相关，尤其与完成度关系较强；

耐用性与完成度负相关；

颜色与情感无关，与完成度负相关。

3.研究应用：优化产品描述和制作工艺，指导颜色选择，加强消费者教育提升满意度。



4 数据分析

XDU

4.4 特征重要性分析

1. 展示特征的重要性排序结果，显示每个**特征对目标变量的影响程度。**

2.情感：得分0.913265，表明情感评价是预测服装质量的主要因素，**影响最为显著。**

3.其他特征：耐用性、构造、材料、颜色和完成度得分均低于0.02，显示这些物理特性对预测的贡献相对较小。

4.建议：电商平台应**重点关注情感分析**，以提高服装评论的预测准确性，同时考虑**优化情感分析技术。**

```
X = df.drop(columns=['Cons_rating', 'Title', 'Review', 'Cloth_class'])
y = df['Cons_rating']
model = RandomForestRegressor()
model.fit(X, y)
feature_importance = pd.Series(model.feature_importances_, index=X.column

print("\nFeature Importance:")
print(feature_importance)
```

```
Feature Importance:
Sentiment       0.913265
Durability      0.024430
Construction    0.020948
Materials       0.014762
Color           0.013782
Finishing       0.012814
dtype: float64
```

4 数据分析

XDU

4.5 LDA主题分析

内容描述： LDA主题分析显示电商服装评论的主要关注点是尺寸合身度、质量舒适度和款式颜色。

```
# 找出每个主题的关键词
feature_names = vectorizer.get_feature_names_out()
top_words = []

for topic_idx, topic in enumerate(lda.components_):
    top_words_idx = topic.argsort()[:-10-1:-1]
    top_words.append([feature_names[i] for i in top_words_idx])

# 打印出每个主题的关键词
for i, words in enumerate(top_words):
    print(f"Topic {i+1}:")
    print(", ".join(words))
```

分析建议：

1.优化尺寸设计： 增加尺码选项，改善合身度。

2.提升质量： 使用优质面料，改进制造工艺。

3.丰富款式： 提供多样化的款式和颜色选择。

```
Topic 1:
dress, size, small, like, just, love, fabric, fit, large, ordered

Topic 2:
color, sweater, good, jacket, like, great, shirt, quality, nice, fit

Topic 3:
fit, like, size, pants, just, waist, ordered, quality, don, good

Topic 4:
love, comfortable, dress, great, wear, jeans, perfect, soft, fit, flattering

Topic 5:
size, fit, perfect, fits, great, wear, perfectly, bought, long, comfortable
```

5 讨论与结论

本实验采词云图、用情感分析、LDA主题分析相结合的方法，对服装产品类的消费者评论进行了综合分析。

XDU

5

词云图

直观展示服装评论文本中词汇的频率和重要性，
帮助快速识别主要主题和关键词，指导改进。

情感分析

揭示了消费者对服装产品的情感倾向。
大部分评论体现了积极情感，但也发现了改进机会。

LDA主题分析

挖掘了评论中的关键主题和关注点。
指引了产品设计和服务质量的优化方向。

缺失值处理方式不当

实验中处理数值类型数据缺失值的方式是全部置0。
由于部分属性列缺失比例高，填充后0的比重很大。

不涉及异常值处理

本次实验中的数据集含有约350条异常值数据，约占总量的1%。
但异常原因是单元格的不规则错位，难以修复

XDU

产品评论文本挖掘以及综合应用

感谢聆听　　敬请指正

汇报：甘逸民

厚 德　求 真　励 学　笃 行