Advances in Applied Mathematics 应用数学进展, 2021, 10(1), 128-136
Published Online January 2021 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2021.101015

Hans汉斯

基于细粒度情感的文本挖掘及可视化分析

程　斌，高圣国*

上海工程技术大学，上海
Email: cb951013@163.com, *gsg1688@163.com

收稿日期：2020年12月16日；录用日期：2021年1月5日；发布日期：2021年1月20日

摘　要

针对当前文本挖掘情感分析缺乏系统性分析研究，建立综合评价指标，系统的分析产品间优劣势，明确改进方向。构建基于细粒度情感分析的模型，首先通过对评论文本进行预处理及分词，再运用LDA主题模型构建属性词典，运用知网情感词词库结合网络新词构建情感词典；接着建立评论有用性规则与情感打分规则，对有用短语打分，获取情感数据集；最后建立四大评价指标，对三款手机进行综合评价及可视化分析。模型数据结果表明，四大指标能够显著突出产品间优劣势，可以帮助生产者更快更准确的了解重点发展方向，也可以帮助消费者更便利的选择钟爱的产品。

关键词

细粒度情感分析，文本挖掘，在线评论，可视化分析

Text Mining and Visualization Analysis Based on Fine-Grained Sentiment

Bin Cheng, Shengguo Gao

Shanghai University of Engineering Science, Shanghai
Email: cb951013@163.com, *gsg1688@163.com

Received: Dec. 16th, 2020; accepted: Jan. 5th, 2021; published: Jan. 20th, 2021

Abstract

For the current text mining sentiment analysis, there is a lack of systematic analysis and research, this paper establishes comprehensive evaluation indicators, systematically analyzes the advan-

*通讯作者。

文章引用: 程斌, 高圣国. 基于细粒度情感的文本挖掘及可视化分析[J]. 应用数学进展, 2021, 10(1): 128-136.
DOI: 10.12677/aam.2021.101015

tages and disadvantages of products, clarifies the direction of improvement and constructs a model based on fine-grained sentiment analysis. Firstly, the comment text is preprocessed and segmented, and then the LDA topic model is used to build the attribute dictionary, and the HowNet sentiment vocabulary is used to build the sentiment dictionary with new words on the Internet; then, in order to obtain the sentiment data set, use the comment usefulness rule and sentiment score rules, scoring useful phrases; finally, four major evaluation indicators are established to conduct comprehensive evaluation and visual analysis of three mobile phones. The results of the model data show that the four indicators can significantly highlight the advantages and disadvantages of the products, which can help the producers to understand the key development direction more quickly and accurately, and also help consumers to choose the products they like more conveniently.

Keywords

Fine-Grained Sentiment Analysis, Text Mining, Online Review, Visual Analysis

Copyright © 2021 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 20 世纪的结束，电子商务在中国迅速发展，截至 2019 年 6 月，我国网络购物用户规模达 6.39 亿[1]。数量庞大的网络购物用户在电商平台购买商品并对商品做出的评价后，留下了巨量的评论信息，这些评论信息就被学者们称为电商的在线评论[2]。在线评论中的文本数据传达了消费者对于电商平台和商品等多方面的信息，这些信息不仅是其他潜在消费者决策前的重要参考，对于商品生产厂商的产品改进与研发也是十分重要。

本文通过运用 LDA 主题模型与完善补充产品属性及情感词典，构建细粒度情感分析模型，建立基于文本挖掘的评价指标，并通过实证，进行评价、建议与可视化分析。

2. 基于细粒度情感分析的模型构建

在线评论分析研究模型如下图 1 所示。

模型共包含五个步骤，分别是评论数据爬取及预处理[3]，Jieba 分词，运用隐含狄利克雷主题模型(LDA) [4]获取产品属性及属性词[5]，在线评论情感词典构建[6]，有用短语及情感打分，数据分析及可视化。

3. 实证分析

3.1. 实验数据来源

智能手机品牌繁多、功能多样，也是人们生活中广泛使用的一种产品。本文就以智能手机的在线评论作为实验数据，浏览相关测评网站[7] [8]，选取 OPPO Reno 10 倍变焦版、vivo X27 和华为 P30PRO 这三种手机的在线评论为爬取对象，以下简称为 OPPO、vivo 和 HUAWEI。京东商城作为一家较为主流的电商平台，受到不少消费者青睐，通过在京东商城网页搜索上述三款手机，编写 Python 爬虫程序，爬取在线评论。数据爬取时间是 2018 年 7 月 22 日。爬取结果：OPPO 在线评论 2954 条，vivo 在线评论 3946 条，HUAWEI 在线评论 3870 条。

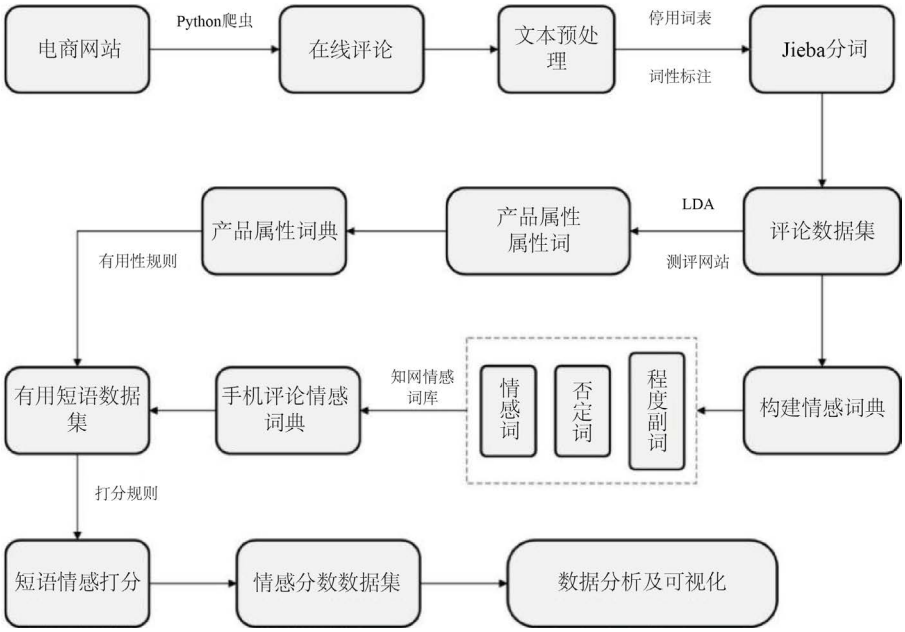

Figure 1. Online comment sentiment analysis model
图 1. 在线评论情感分析模型

3.2. 预处理及 Jieba 分词

为了提高数据的有效性，对爬取的评论数据进行以下预处理工作：① 删除评论为“此用户没有填写评论!”的评论；② 删除重复评论；③ 删除明显带有广告色彩的评论[9]。预处理后，三款手机评论数量分别为：2831 条、3730 条和 3686 条。

通过调用 Python 中的 Jieba 分词，对预处理后的数据进行分词与词性标注，每条评论分词结果按行保存到 txt 中。文本中包含了多种词性，如：名词、代词、动词、形容词、量词、副词、助词、连词等等，其中助词、连词等词无实际意义，影响数据处理与程序运行，因此先在 Python 中对评论文本进行词性标注，再通过加载停用词表将这些无用词与标点符号删除。本文采用的停用词表，是四川大学机器智能实验室停用词库、哈工大停用词表和百度停用词表汇总、去重后的得到的停用词表。

3.3. LDA 确定产品属性及属性词

在 Python 中按行读取分词后的评论，调用第三方模块 LDA 包，通过设置参数，主题个数(n_topics)和迭代次数(n_iter)，训练模型，对评论文本进行产品属性及属性词提取。经过多次调试后，在 n_topics = 6，n_iter = 10000 时，输出每个主题中前 15 个词(Top-N = 15)，此时模型结果较优，得到 6 个产品属性，90 个属性词。通过参考中关村在线的手机板块、主流手机厂商官网和其他学者关于手机属性的相关文献，对初次结果进行调整，最后得到 6 个产品属性和 87 个属性词。如表 1。

3.4. 手机评论情感词典构建

正如上文中所提到的，现有的情感词典不能满足对于在线评论文本分析的要求。经过词性标注与分词后，对分词结果进行整理，筛选出特有的词汇与网络新词，并结合已有情感词典，最终得到基于手机在线评论的情感词典。如表 2 所示。接着，根据每个词所带代表的含义与强度，为情感词典中的词设置权重，情感词的权重取值分别是：−1 和 1，程度副词的权重取值分别是：0.8，1.2，1.6 和 2，否定词的权重为−1。

Table 1. Mobile phone attribute vocabulary
表 1. 手机属性词表

属性	属性词举例
外形外观	包装，颜色，做工，颜值，外观，配色，样式，机身，设计
摄像拍照	拍照，像素，摄像头，夜景拍照，相机，变焦，摄影，自拍
电池续航	续航，充电速度，电池，耗电，续航能力，耗电量，待机时间
屏幕显示	屏幕，屏占比，显示，分辨率，画质，画面，视觉
系统性能	游戏，性能，系统，运行速度，cpu，处理器，内存，反应速度
体验服务	指纹解锁，刷脸，音质，震动，性价比，散热，重量，手感

Table 2. Emotional dictionary
表 2. 情感词典

词性		举例
否定词		不，无，勿，非，未，不能，没有
情感词	积极	不错，完美，出色，爽，棒，给力，优秀，惊喜，666
	消极	丑，差，迟钝，瑕疵，迟钝，简陋，卡顿，积灰
程度副词	等级 1	非常，极其，极，极度，超级
	等级 2	很，很是，特，太，相当，挺
	等级 3	还，较，较为，比较
	等级 4	略，稍微，有点，偏

3.5. 有用性评论

以属性词典和手机评论情感词典作为依据，本文进一步规定，一个短句中同时包含属性词和情感词时，这个短句被认为是有用的。[10]如下公式：

$$\beta_i = \beta_i^s \times \beta_i^q \tag{1}$$

其中，$\beta_i^s = \{0,1\}$，$\beta_i^q = \{0,1\}$；当 $\beta_i^s = 1$ 时，表示第 i 个短句中包含属性词；$\beta_i^s = 0$ 时，表示第 i 个短句中不包含属性词。当 $\beta_i^q = 1$ 时，表示第 i 个短句中包含情感词；$\beta_i^q = 0$ 时，表示第 i 个短句中不包含情感词。当 $\beta_i = 1$ 时，表示第 i 个短句是有用的；$\beta_i = 0$ 时，表示第 i 个短句是无用的。

以公式(1)为基础，通过 Python 代码遍历每个短句、属性词和情感词，计算短句有用性，将 β_i 值为 1 的短句保存为列表。

3.6. 基于有用短语的情感打分

有用短语包含属性词和对属性的情感描述，其中情感描述必定包含情感词，可能含有否定词和程度副词。基于情感词、否定词和程度副词间的不同组合，本文列举短句中情感描述的 5 种情况作为情感打分依据。如表 3。

Table 3. Emotional scoring formula
表 3. 情感打分公式

序号	情景	举例	公式
1	情感词	好看	$Score = S_{情感词}$
2	程度副词 + 情感词	很好看	$Score = S_{程度副词} * S_{情感词}$

Continued

3	否定词 + 情感词	不好看	$Score=(-1)*S_{情感词}$
4	程度副词 + 否定词 + 情感词	很不好看	$Score=(-1)*S_{程度副词}*S_{情感词}$
5	否定词 + 程度副词 + 情感词	不是很好看	$Score=(-0.2)*S_{程度副词}*S_{情感词}$

表 3 中，$Score$ 为短句的情感分数，$S_{情感词}$ 和 $S_{程度副词}$ 分别为情感词和程度副词的权重值，在 3 和 4 两种情况下，否定词权重为−1；在第 5 种情况中，否定词权重取−0.2。

基于以上公式，通过 Python 编写的代码读取列表中所有短句并进行打分，以分值与短句一一对应的形式保存为 excel 文档，此时得到对于手机属性的全部情感分值，则该属性的情感分值为：

$$S(N_i)=\sum_{j=1}^{n}S(w_{ij}) \tag{2}$$

式中：$S(w_{in})$ 为属性词情感分值；$S(N_i)$ 为第 i 个属性的情感分值。

3.7. 基于情感打分的评价指标

为了比较三款手机属性整体情况，依据情感打分获得的数据，建立以下评价指标：满意度 SA、关注度 AT、待改进程度 IM 和方差 VA。

根据公式(2)，当正向分值占比越高，则表明消费者的满意度越高，满意度 SA 计算公式如(3)：

$$SA=\frac{S(N_i^+)}{S(N_i^+)+\left|S(N_i^-)\right|} \tag{3}$$

式中：SA 为对属性的满意度；$S(N_i^+)$ 为属性正向分值；$S(N_i^-)$ 为属性负向分值。

消费者对于某一属性提及次数越多，则表明消费者对该属性的关注度越高。当某一属性的关注度越高，且分值为负的评价占比越高，说明该属性待改进程度越高，则关注度计算公式如(4)，待改进程度计算公式如(5)：

$$AT=\frac{T_{N_i}}{NUM(T_N)} \tag{4}$$

$$IM=AT\times\frac{T_{N_i}^-}{T_{N_i}}=\frac{T_{N_i}^-}{NUM(T_N)} \tag{5}$$

式中：AT 为属性的关注度；IM 为属性的待改进程度；T_{N_i} 为某属性提及的次数；$T_{N_i}^-$ 为属性评价分值为负的次数；$NUM(T_N)$ 为所有属性出现的总次数。

某个属性情感分值的方差大小，表明了消费者对该属性情感的离散程度。方差越大，离散程度越高，说明消费者对于该属性的评价分歧较大；方差越小，离散程度越低，说明消费者对于该属性评价较为一致，均值及方差公式如(6) (7)：

$$M_i=\frac{S(N_i)}{T_{N_i}} \tag{6}$$

$$VA=\sum_{i}^{n}\frac{\left(S_{w_{ij}}-M_i\right)^2}{T_{N_i}-1} \tag{7}$$

式中：M_i 为第 i 个属性的情感均值；VA 为情感分值方差；$S_{w_{ij}}$ 为每个属性词的情感分值；T_{N_i} 为第 i 个属性的提及次数。

依据上述公式，对情感得分进行整理，得到下表 4：

Table 4. Three types of mobile phone evaluation index results
表 4. 三类手机评价指标结果

手机类型	手机属性	满意度 *SA*	关注度 *AT*	待改进程度 *IM*	方差 *VA*
Huawei	外形外观	0.989	0.193	0.002	0.183
	摄像拍照	0.992	0.277	0.002	0.202
	电池续航	0.945	0.083	0.005	0.440
	屏幕显示	0.897	0.068	0.008	0.647
	系统性能	0.929	0.226	0.019	0.496
	体验服务	0.908	0.153	0.016	0.674
vivo	外形外观	0.979	0.197	0.005	0.246
	摄像拍照	0.989	0.265	0.003	0.224
	电池续航	0.905	0.091	0.009	0.563
	屏幕显示	0.982	0.102	0.002	0.228
	系统性能	0.915	0.192	0.018	0.481
	体验服务	0.777	0.153	0.037	1.082
oppo	外形外观	0.939	0.129	0.009	0.466
	摄像拍照	0.951	0.234	0.014	0.462
	电池续航	0.831	0.095	0.016	0.860
	屏幕显示	0.961	0.115	0.005	0.386
	系统性能	0.855	0.209	0.028	0.737
	体验服务	0.498	0.218	0.106	1.478

3.8. 数据分析与讨论

为了直观了解消费者对三类手机产品的满意度与情感离散程度，本文依据表 4 中的满意度与方差数据，建立图表如图 2 和图 3 所示。

从图 2 整体来看，消费者对于三款手机的外形外观和摄像拍照属性都有很高的满意度，对于电池续航、屏幕显示和系统性能满意度较高，而对于体验服务的满意度整体偏低。细分看来，华为款手机在外形外观、摄像拍照、电池续航、系统性能和体验服务上，消费者满意度领先于另外两款手机，屏幕显示满意度低于 vivo 款手机和 oppo 款手机，vivo 款手机屏幕显示满意度最高，而在其他方面，vivo 款手机满意度都处于中间位置，oppo 款手机满意度处于末尾。

结合图 3 情感方差，可以看出三款手机满意度越高，情感方差越小，说明消费者情感倾向越一致。总体看来，华为款手机方差最小，且满意度最高，vivo 款手机次之，oppo 款手机最后。三款手机体验服务属性的方差都明显高于其他属性，反映了在体验服务方面存在较大的问题，值得注意。

随着科技的快速发展和手机的广泛使用，手机越来越被当作生活中社交与娱乐的重要工具。本文依据表 2 数据建立关注度与待改进程度图表，如图 4 和图 5 所示。

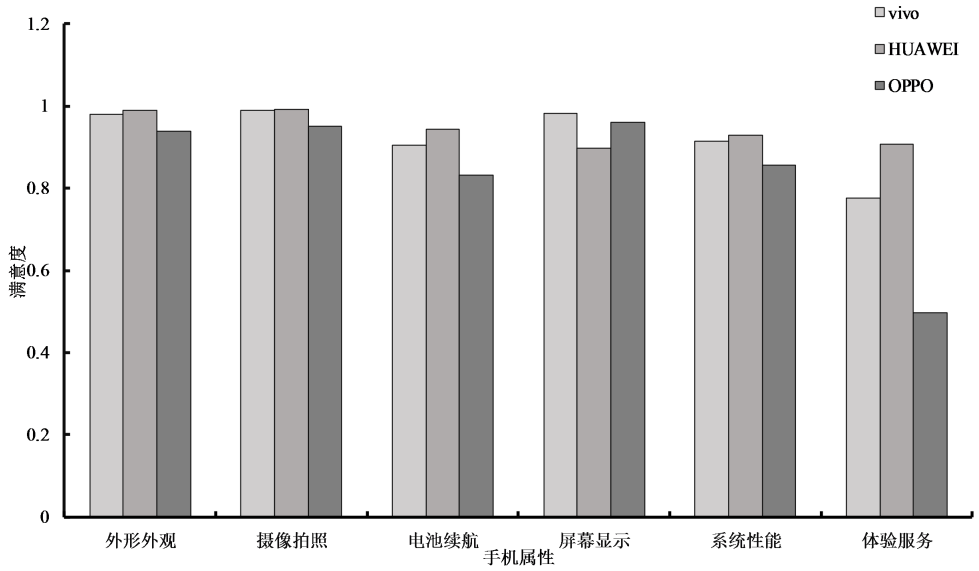

Figure 2. Histogram of satisfaction
图 2. 满意度柱状图

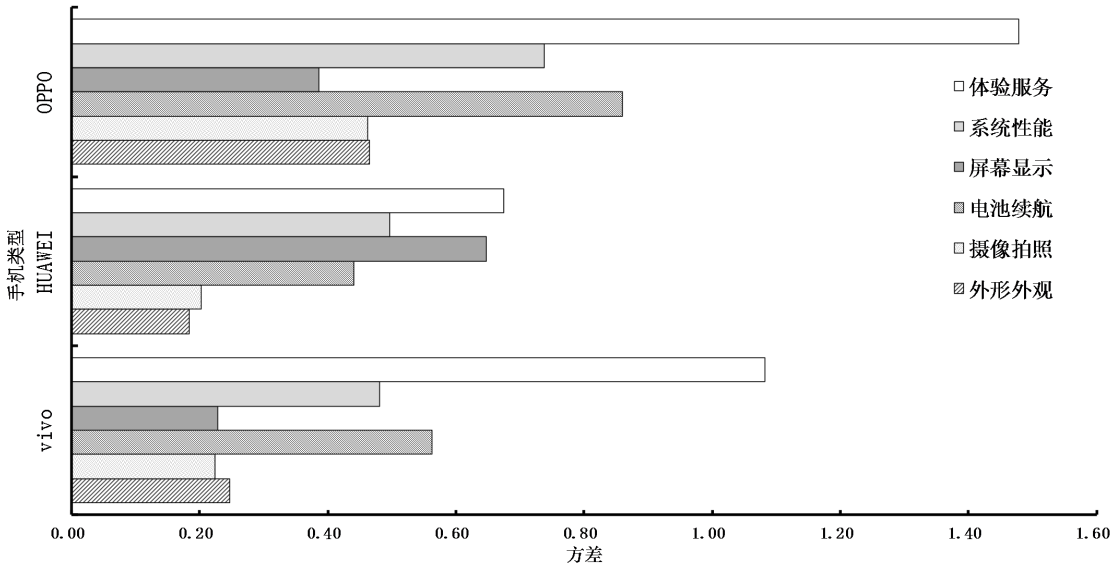

Figure 3. Bar chart of variance
图 3. 方差条形图

从图 4 可以看出，消费者对于手机产品属性的关注度前三由高到低，依次是：拍照摄像、系统性能、体验服务。手机作为记录日常生活和出门旅行的重要工具，手机的拍照摄像越来越得到人们的重视，各大手机厂商也瞄准了这一潮流，向优化升级拍照摄像功能靠拢。其次就是手机的系统性能，随着电竞产业的蓬勃发展，网络游戏也逐渐从端游向手机游戏发展，这就对手机的系统性能提出了较高的要求，一个优越的系统性能能够带来完美的游戏体验，这也是各大厂商研发手机时所注重的。体验服务这一属性是综合手机其他属性，消费者使用过程中所感受到的优劣，根据图 2 和图 3 信息，三款手机在体验服务方面均获得较低的满意度，且方差较大，根据关于体验服务属性的评论文本构建词云图，如下图 6。

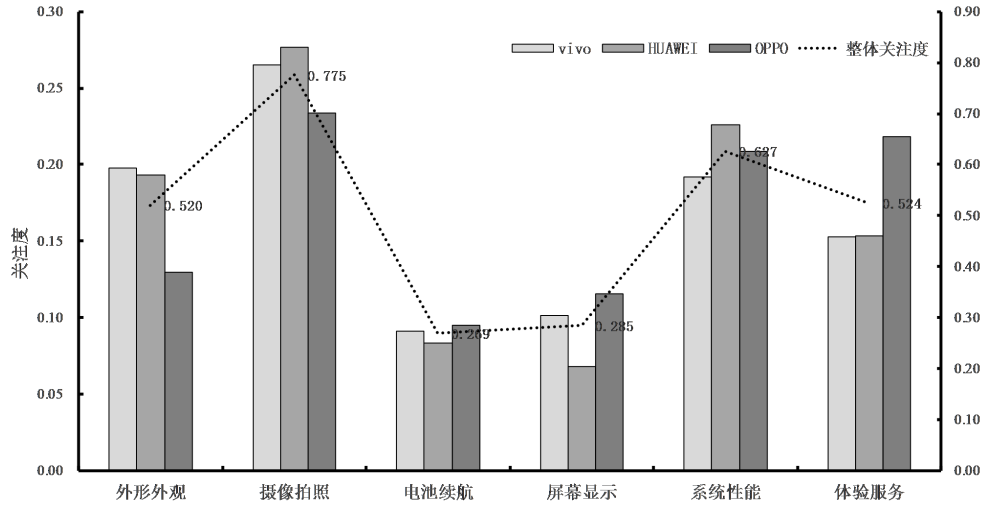

Figure 4. Histogram of attention
图 4. 关注度柱状图

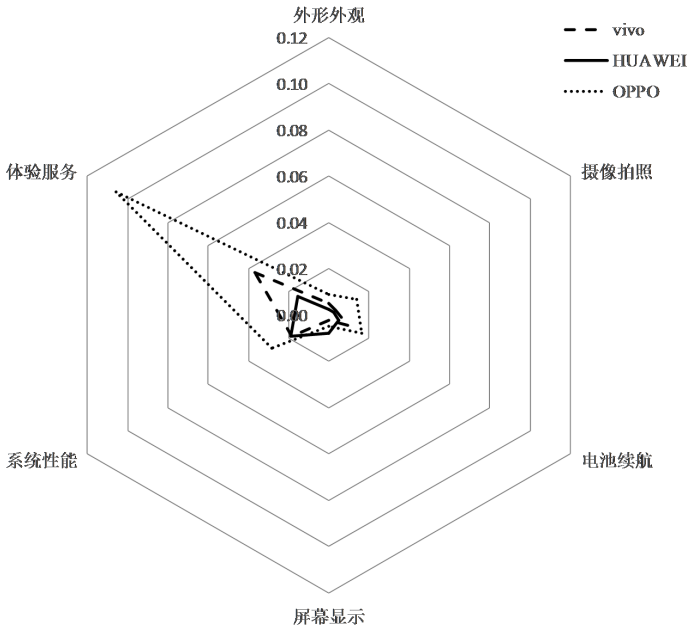

Figure 5. Radar chart to be improved
图 5. 待改进程度雷达图

Figure 6. Word cloud
图 6. 词云图

结合图 5 与图 6，可以看出体验服务方面属性明显急需改进提高，且消费者反馈信息中尤其指出手机的厚度、重量等方面的使用体验满意度很低。其次在系统性能属性方面，也表现出了较高的待改进程度，多体现在运行速度与游戏优化方面。

综合以上信息，提出以下建议：

1) 外形外观、摄像拍照和屏幕显示需稳中求进。总体看来，虽然三款手机在外形外观、摄像拍照和屏幕显示方面，都获得一致的高满意度，但仍然得到消费者的持续关注，分别排在第一、第四和第五。因此，在原有基础上，坚持对这些方面的研发，关注消费者消费趋势变化，增强客户粘性，维持并扩大客户群体。

2) 手机游戏产业以及娱乐类应用软件的迅速发展对手机的电池续航与系统性能提出了更高的要求。根据 Niko Partners 与 QuestMobile 发布的调查报告显示，中国 2018 年手机游戏行业总收入达到 156.3 亿美元，同比上升 28.9%；中国互联网用户每天在手机上的娱乐时间平均达到 4.7 小时。长时间的手机使用与高负荷的游戏过程需要更持久的电池续航与强悍的系统性能。因此，在这两个方面需要更快的迭代更新，才能获得消费者青睐。

3) 用户体验是重中之重。三款手机在体验服务方面都获得最低的满意度与最大的方差，消费者反应在手机重量、厚度、散热和音质等方面使用体验很差。因此要重点把握手机尺寸大小，减轻手机重量，减小手机厚度，贴近大众使用习惯；优化散热系统设计，保证使用手感。

4. 结束语

手机产品种类繁多、更新换代快，如何不断地获取消费者多方面的反馈，并从繁杂的信息中获得有用数据，分析产品优势劣势，是手机产品正确改进的重点。本文通过对爬虫获取的手机在线评论数据进行细粒度情感分析，建立四大评价指标，进行数据可视化，直观了解消费者情感趋势与产品优劣势，并给出相关建议，指明正确的改进方向与策略。

参考文献

[1] 中国互联网络信息中心. 第 44 次《中国互联网络发展状况统计报告》[EB/OL]. http://www.cac.gov.cn/2019-08/30/c_1124938750.htm, 2019-8-30.

[2] 张玉峰, 朱莹. 基于 Web 文本挖掘的企业竞争情报获取方法研究[J]. 情报理论与实践, 2006(5): 563-566.

[3] 薛为民, 陆玉昌. 文本挖掘技术研究[J]. 北京联合大学学报(自然科学版), 2005(4): 59-63.

[4] 张振华, 许柏鸣. 基于在线评论文本挖掘的商业竞争情报分析模型构建及应用[J]. 情报科学, 2019, 37(2): 149-153+160.

[5] 王克勤, 毋凤君. 面向产品设计改进的在线评论挖掘[J]. 计算机工程与应用, 2019, 55(19): 235-245+252.

[6] 林崇德, 杨治良, 黄希庭. 心理学大辞典[M]. 上海: 上海教育出版社, 2003.

[7] Wilson, T., Wiebe, J. and Hoffmann, P. (2005) Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis, Vancouver, British Columbia, Canada, F oct, 2005. Association for Computational Linguistics.

[8] 唐晓波, 刘广超. 细粒度情感分析研究综述[J]. 图书情报工作, 2017, 61(5): 132-140.

[9] 杨东红, 吴邦安, 陈天鹏, 等. 基于京东商城评价数据的在线商品好评、中评、差评比较研究[J]. 情报科学, 2019, 37(2): 125-132.

[10] 杨程, 谭昆, 俞春阳. 基于评论大数据的手机产品改进[J/OL]. 计算机集成制造系统, 1-19. http://kns.cnki.net/kcms/detail/11.5946.TP.20190606.1053.008.html, 2020-10-03.