基于情感分析混合模型的用户评论主题分析

——以 vivo 手机为例

张倩男

(黄河交通学院 基础教学部，河南 焦作 454950)

摘要：为了帮助商家了解用户需求和产品问题，进一步提升产品和服务质量，对 vivo 手机评论数据进行基于关键词的词云图分析、社会语义网络分析、舆情时间序列可视化分析，挖掘用户关注焦点与手机特征的内在联系和用户情感倾向趋势；然后对评论数据进行基于 LDA 的主题特征分析，继而提出一种基于 Word2vec 和 SVM、LDA 的混合算法模型，挖掘用户正向和负向情感评论的潜在主题，得到不同情感倾向下用户对 vivo 手机不同方面的反映情况。分析结果表明，基于混合算法的挖掘结果比基于关键词的可视化分析、基于 LDA 的主题分析更清晰，更具准确性，为商家提供的建议更有意义。

关键词：数据挖掘；词云图；主题分析；情感分析；Word2vec

中图分类号：TP391 **文献标志码**：A **文章编号**：1671－1807(2022)04－0347－08

伴随着全球进入智能手机时代，手机产品不断丰富，产品竞争也日益加剧。随着电子商务的发展，越来越多的消费者选择通过网络平台购买手机，并在平台发表产品评论。评论包含产品不同属性的评价、整体性评价以及与其他产品的对比评价等信息。挖掘在线评论中蕴涵的潜在信息，能有效帮助商家实现自身产品与服务的优化，进行营销与竞争策略调整，完成精细化管理，进一步提升企业竞争力，同时也能帮助消费者做出更加明智的消费决策。

主题模型作为可以细粒度挖掘文档主题和情感分布的无监督模型，许多学者将其引入到情感分析研究中。陈晓美和关心惠[1]在 LDA 在线舆情视图提取的基础上，结合舆情主题和情感因素对网络评论提取了主要观点。万晓霞[2]提出了一种改进的 LDA 建模方法，利用 TF-IDF 值对文本词的权重进行过滤，提高了热门话题发现的速度和准确性。Hu 等[3]运用 LDA 模型对时事新闻的社交媒体评论数据进行分析，得到用户的意见。田贤忠等[4]基于 BBS-LDA 进行了论坛主题的挖掘。曾寰等[5]基于语义相似度对商品评论进行 LDA 主题情感分类研究。

在文本分类的研究中，谢宗彦等[6]基于 Word2vec 为酒店在线评论构建了一个情感分析的模型，取得较好的效果。吴龙峰[7]提出了一种结合神经网络语言模型 Word2vec 和文档主题模型 LDA 的文本特征表示模型。Zhang 等[8]为了得到语义特征，提出了一种基于 Word2vec 和支持向量机性能的情感分类方法。文献[9-11]也分别基于 Word2vec 对情感分类进行了研究。Sharma 等[12]从预先训练好的 word2vec 模型中生成词向量，并利用 CNN 层提取出更好的特征用于短句分类。

为了更有效挖掘用户评论的语义信息，本文以 vivo 手机用户评论数据为研究对象，对用户评论进行可视化分析和主题模型分析。进行词频统计，并绘制词云图，挖掘用户对 vivo 手机的关注焦点；进行社会语义网络的可视化分析，挖掘手机评论特征的内在联系；使用 SnowNLP 处理用户评论信息，在时间轴上观察用户在特定时间段内的情感倾向趋势，定位用户负面评论信息。基于整体数据集进行 LDA 主题特征分析，挖掘用户主要讨论话题；为提高主题分析在不同情感倾向下热门关注点反映情况的精确度，将 LDA 和基于 Word2vec 的 SVM 算法结合，分别挖掘用户正向和负向情感评论的潜在主题，得到不同情感倾向下用户对 vivo 手机不同方面的反映情况。

1 数据来源及处理

根据市场调研机构 Canalys 发布的《2019 年中

收稿日期：2021-11-29

作者简介：张倩男(1994—)，女，河南商丘人，黄河交通学院基础教学部，助教，硕士，研究方向为机器学习、情感分析。

国大陆智能手机出货量及市场份额》报告，发现2019 年在中国大陆市场 vivo 手机出货量虽然排名第二，但同比表现下滑趋势，故本文选取 vivo 手机的用户评论作为研究对象，选定网络爬虫工具——八爪鱼采集器，通过模仿用户的网页操作，指定数据采集逻辑和选择采集的数据，进行数据采集的流程设计，完成采集规则的制定，然后基于流程设计进行用户评价界面相关信息的采集，最终共采集到近 3 万条 vivo 手机用户评论数据，采集字段包含用户 id、用户评分、评价内容、手机型号、购买时间。

在分析之前，需要通过数据清洗完成数据的规整，以提高后续情感分析的精确性。文本评论的处理主要包括：

1）初步清洗。通过定位、筛选、查找、排序等功能对原始数据进行简单的预处理，如删除卖家回应评论部分以及无实质评论内容部分。

2）文本去重。采用比较删除法，去除文本评论数据中无用的自动评论、重复评论以及抄袭的评论内容，即对完全重复的语料进行两两对比，仅保留一条有用的文本评论信息，删除其他重复评论，确保数据的唯一性。

3）机械压缩去重。由于数据量较大，且文本数据质量良莠不齐，包含很多没有意义的文本数据，故需要对其进行压缩，去掉连续重复的无意义词汇。评论压缩语句效果对比结果见表 1。

表 1　用户评论语句压缩前后对比结果

原语句	被压缩语句
手机外观真的非常非常非常的好看！	手机外观真的非常好看！
物美价廉！最重要的是性价比很高，值得购买，赞赞赞赞赞！	物美价廉！最重要的是性价比很高，值得购买，赞！
不错不错不错，拍照非常清晰，运行速度快，物超所值，值得购买	不错，拍照非常清晰，运行速度快，物超所值，值得购买

4）中文分词。jieba 中文分词使用基于统计的分词方法，基于前缀词典实现对所有词汇的扫描，然后将一条语句中所有可能的生成词汇构成有向无环图（DAG），基于 DAG 图，采用动态规划计算最大概率路径找出最大切分组合。jieba 中文分词的精确模式比较适合文本分析，能够将句子最精确地切开。本文数据是用户在线评论文本，故采用结巴分词的精确模式进行分词。

5）去停用词。去停用词的目的是为了减少信息冗余，提高分析的效率和准确性，而去停用词的关键在于停用词表的维护。本文使用“哈工大停用词词库”“四川大学机器学习智能实验室停用词库”“百度停用词表”3 种停用词库，对停用词人工整理、匹配、筛选、去重；利用 Python 语言筛选对手机评论数据无帮助和无意义的词汇，加入停用词词典，停用词表共包含 2 185 个词汇；最后利用 Python 语言基于新的停用词表对分词后的用户评论数据进行二次过滤，实验效果显著，实验结果见表 2。

表 2　去停用词结果

分词结果	去停用词结果
运行速度比较流畅，还是不错的	运行速度流畅不错
拍照效果蛮好的，成像很清晰，比较喜欢前置的自拍效果	拍照效果好成像清晰喜欢前置自拍效果
不错，外观很漂亮，电池非常耐用	不错外观漂亮电池耐用

2　用户评论的可视化分析

2.1　基于 TF-IDF 的文本关键词抽取

TF-IDF 是一种衡量文档中某个词对该篇文档重要程度的计算方法，一个词语在一篇文章中出现次数越多，同时在所有文档中出现次数越少，越能够代表该文章。文本关键词抽取流程如图 1 所示。

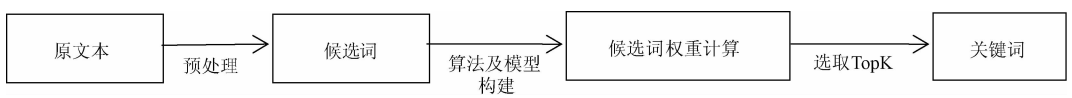


图 1　文本关键词抽取流程

基于处理之后的数据，采用 TF-IDF 算法处理文档词项，获得更合理的更能代表这篇文档特点的向量，在转化成文档向量后，依据权值大小进行关键词提取，从而进行不同文档间的相似度分析。TF-IDF 公式为

$$\text{TF-IDF} = \text{TF} \times \text{IDF} = \frac{N_A}{N} \times \log \frac{|D|}{|D_A|+1} \tag{1}$$

式中：TF 为指词在文章中出现的次数，即词频；IDF 为衡量词的常见程度，即逆文档频率；N_A 为该文档词项 A 的总数；N 为该文档总词数；$|D_A|$ 是包含词项 A 的文档数；$|D|$ 是语料库中的文档总数。通过 TF-IDF 公式，可以计算出特定词对于表现这篇文档主题的贡献度。

2.2 评论数据可视化分析

2.2.1 词云图分析

在用户评论的焦点分析中，首先基于词法分析做评论的分词和词条的词性标注，文本过滤筛选符合关键词搜索域的词条；继而基于 TF-IDF 算法实现关键词的获取，提取出的关键词浓缩了用户评论中的精华信息，能反映出用户的关注点、情绪和认知，产品的潜在竞争力等信息；之后对关键词进行词频统计，提取与产品内容、属性有关的关键词；最后对前 101 个关键词基于词云图展示评论热点与焦点。词云图如图 2 所示。

图 2 词云图

通过结合词云图和词频统计结果可以看出，除了表示研究对象的“手机”外，“不错”“喜欢”“满意”是评论中较为突出的高频词汇，其均与用户态度有关，代表大部分用户的总体态度是较正面的。与手机性能特征相关的词汇有“流畅”“运行”“系统”“性能”“处理器”“配置”等，这些词出现频率也较高。“屏幕”“漂亮”“外观”“好看”“颜色”等反映手机外观的词汇，说明用户对手机外观比较关注。“拍照”“清晰”“照相”“摄像头”高频词说明用户对手机拍照功能也比较关注。“快递”“服务”“物流”“态度”反映用户购物体验特征的词汇出现频次也不低，表明用户对购物过程中的购物体验有着较高的要求。“电池”“耐用”“电量”表明有些用户关注手机的续航能力。“发热”“不好”消极词汇的出现说明用户对手机某些体验有所不满。

总体来看，用户对 vivo 手机的关注点主要集中在手机的性能、外观、拍照功能、续航能力，另外用户也比较关注购物体验过程，但是对这些关注点的态度并不能在词云图中体现，需要进一步研究。

2.2.2 社会语义网络分析

采用 ROSTCM6 的语义分析工具进行社会网络和语义网络分析，生成社会语义网络结构图，以图形化的方式揭示词与词之间的结构关系，对用户评论文本数据集进行进一步的关联分析，挖掘潜在信息。社会语义网络图如图 3 所示。

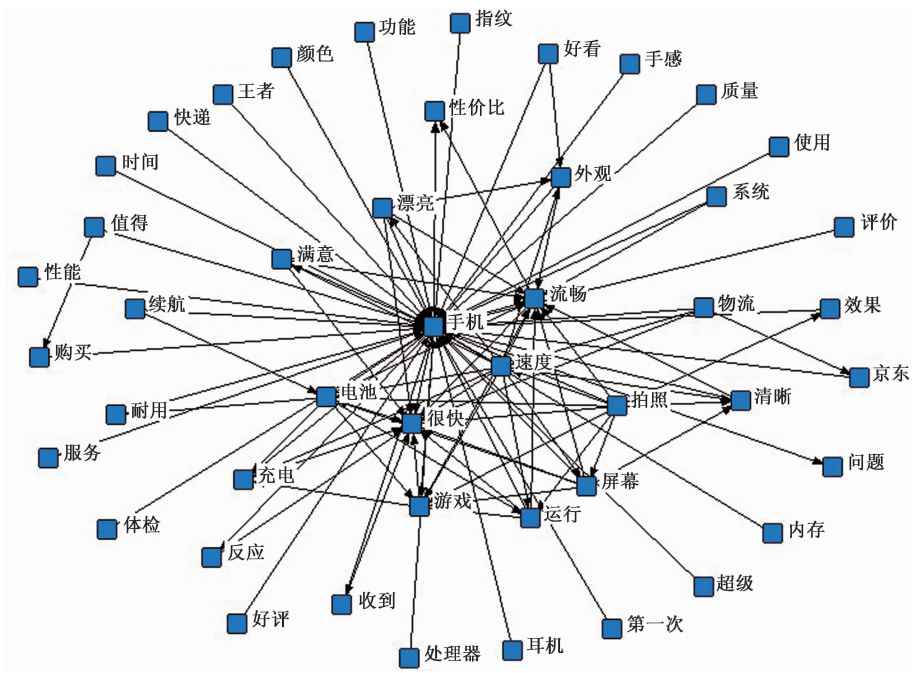


图 3 用户评论社会语义网络图

社会语义网络分析步骤如下：

1)将清洗处理完毕文本的数据导入 ROSTCM6 提取高频词。

2)根据自定义的过滤词表(停用词表)过滤无意义的词，形成高频词表。

3)提取行特征词表，构建高频词和行特征词-共现矩阵词表，构建语义网络图。

通过分析，获得以下发现：

1)结合语义网络关系词频统计结果和语义网络图进行分析，社会语义网络图以“手机”为核心节点，主要表现 vivo 手机系统、电池属性、拍照方面的功能性信息，另外很快、流畅、满意、漂亮等词表明用户对 vivo 手机评价较为积极。

2)次级节点基本以核心节点为中心向周围辐射分布，但其中也存在局部的簇群关系，揭示出主要问题之间的潜在关联：主要表现手机的运行速度快、拍照速度快、充电速度快、玩游戏速度快，同时用户对物流速度比较满意。

3)将“流畅”作为三级节点。主要表现手机在玩游戏、运行、系统、拍照方面比较流畅不卡顿。

4)其他节点。与“外观”相关的漂亮、好看等词表现用户对手机外观比较满意；与“电池”相关的续航、耐用表明手机电池续航时间长。

2.2.3　舆情时间序列可视化分析

情感分析的目的是为了找出说话者/作者在某些话题上或者针对一个文本两极的观点的态度。利用 SnowNLP 情感分析工具处理用户评论信息，其返回值为正面情绪的概率，越接近于 1 表示正面情绪，越接近于 0 表示负面情绪，纵坐标数值越低代表用户评价情感分析的数值越低。将情感分析的结果在时间轴上以可视化形式呈现出来，展示基于时间轴的信息流，如此便可以直观观测到某一段时间内用户对手机的情感倾向趋势，然后基于用户评论的情感极性定位那些可能有问题的异常点，直观查看这些异常点出现在什么时间，以及它们的数值究竟有多低。从而从这些负面评价出发，针对用户的关注焦点进行挖掘，提取有价值的信息，用于产品的改进和相应的销售政策的制定，对商家具有非常重要的意义。

将全部数据的情感分析图进行展示，从整体上把握用户对该产品的情感倾向。由于本文数据量较大，故最终形成的时间序列图高度集中，数据分布较为密集，从图 4 可以看到，数据集高度集中在图形上方，故用户对手机的总体评价是正面的，有些正面评价情感分析数值极端的高，但是也清晰地发现了许多数值极低的点，这些点对应评论的情感分析数值接近于 0，因此被判定为基本上没有正面情感，该部分评论包含用户对手机各方面的负面评论，具有很高的研究价值。

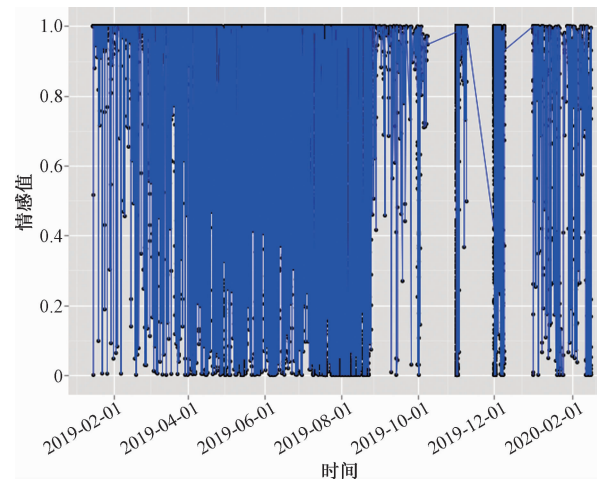


图 4　用户评论的时间序列图

为了清晰地进行舆情分析，抽取 2019 年 12 月份用户评论进行舆情时间序列可视化，如图 5 所示。从时间上看，几乎每隔几天就会出现一次较严重的负面评价(情感值为 0)，因此利用 Python 数据框 Pandas 提供的排序功能找到所有评论里某段时间内情感分析数值较低的评论。将该部分评论使用 TF-IDF 方式提取关键词和权重，发现 2019 年 12 月份的负面评价主要针对客服态度、充电发热问题。针对京东客服服务问题，建议京东平台对客服人员进行素质培训，提高服务质量；针对手机充电发热问题，建议手机制造商对手机电池进行检测，在保证其他优势的基础上，改进手机质量。

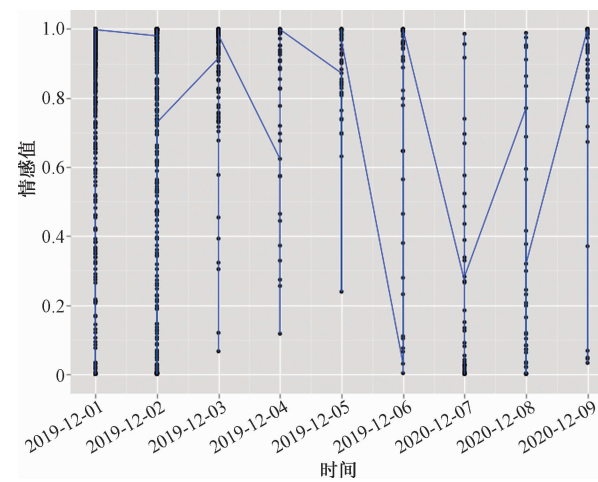


图 5　2019 年 12 月用户评论的时间序列图

3 基于 LDA 的文本主题模型分析

3.1 LDA 主题模型

3.1.1 LDA 主题模型介绍

LDA 是由 Blei 于 2003 年提出的三层贝叶斯概率模型,通过无监督的学习方法发现文本中隐含的主题信息,目的是要以无指导学习的方法从文本中发现隐含的语义维度,包括文档(d)、主题(z)、词(w)三层结构,能够有效地对文本进行建模,挖掘数据集中的潜在主题,进而分析数据集中的集中关注点及其相关特征词。该模型采用词袋的方法对主题词汇进行处理,将一个文档识别成一个词频向量,将文字信息转化成数学信息,定义词表大小为 L,一个 L 维向量(1,0,0,…,0,0)表示一个词,由 N 个词构成的评论即为 $d=(w_1,w_2,\cdots,w_N)$。若商品的评论集 D 由 M 篇评论构成,记为 $D=(d_1,d_2,\cdots,d_M)$,M 篇评论分布着 K 个主题,记为 $z_i(i=1,2,\cdots,K)$。记 α 和 β 为狄利克雷函数的先验参数,θ 为主题在文档中的多项分布的参数,其服从超参数为 α 的狄利克雷先验分布,ϕ 为词在主题中的多项分布的参数,其服从超参数为 β 的狄利克雷先验分布。

LDA 模型假定每篇评论由各个主题按一定比例随机混合而成,混合比例服从多项分布,记为

$$Z \mid \theta = \text{Multinomial}(\theta) \tag{2}$$

而每个主题由词汇表中的各个词语按一定比例混合而成,混合比例也服从多项分布,即为

$$W \mid Z,\phi = \text{Multinomial}(\phi) \tag{3}$$

在评论 d_j 条件下生成词 w_i 的概率表示为

$$P(w_i \mid d_j) = \sum_{i=1}^{K} P(w_i \mid z=s) \times P(z=s \mid d_j) \tag{4}$$

式中:$P(w_i \mid z=s)$ 表示词 w_i 属于第 s 个主题的概率;$P(z=s \mid d_j)$ 表示第 s 个主题在评论 d_j 中的概率。

3.1.2 LDA 主题模型估计

LDA 模型利用吉布斯抽样对参数进行估计,依据为

$$P(z_i=s \mid Z_{-i},W) \propto \frac{(n_{s,-i}+\beta_i)(n_{s,-j}+\alpha_s)}{\left(\sum_{i=1}^{V} n_{s,-i}+\beta_i\right)} \tag{5}$$

式中:$z_i=s$ 表示词 w_i 属于第 s 个主题的概率;Z_{-i} 表示其他所有词的概率;$n_{s,-i}$ 表示不包含当前词 w_i 的被分配到当前主题 z_s 下的个数,$n_{s,-j}$ 表示不包含当前文档 d_j 的被分配到当前主题 z_s 下的个数。

进而得到词 w_i 在主题 z_s 中的分布的参数估计 $\phi_{s,i}$ 和主题 z_s 在评论 d_j 中的多项分布的参数估计 $\theta_{j,s}$,即

$$\phi_{s,i} = \frac{(n_{s,i}+\beta_i)}{\left(\sum_{i=1}^{V} n_{s,i}+\beta_i\right)} \tag{6}$$

$$\phi_{j,s} = \frac{(n_{j,s}+\alpha_s)}{\left(\sum_{s=1}^{K} n_{j,s}+\alpha_s\right)} \tag{7}$$

式中:$n_{s,i}$ 表示词 w_i 在主题 z_s 中出现的次数;$n_{j,s}$ 表示文档 d_j 中包含主题 z_s 的个数。

3.1.3 LDA 的困惑度

对于一篇文章所训练出来的模型对文档属于哪个主题的不确定程度称困惑度,困惑度越低,聚类的效果越好。本文中采用困惑度(Perplexity)确定 LDA 主题模型的最优主题个数,困惑度公式为

$$\text{perplexity}(D) = \exp\left\{-\frac{\sum \log\, p(w)}{\sum_{d=1}^{M} N_d}\right\} \tag{8}$$

式中:$\sum_{d=1}^{M} N_d$ 为测试集的总长度;$p(w)=p(z \mid d)\times p(w \mid z)$ 是测试集中每个单词出现的概率。

3.2 主题模型结果分析

3.2.1 基于 LDA 的主题分析

用户评论整体数据集 LDA 主题提取步骤如下:

1)读取数据,加载自定义停用词表,对数据进行预处理操作,分词、词性标注、去停用词、词和词性构成一个元组。

2)进行特征关键词的限定,由于用户评论数据包含大量的词汇,若考虑全部词汇,一方面将导致数据处理时间过长,另一方面一些不常用的词汇对主题抽取意义不大,故限定从评论文本中提取 5 000 个最重要的特征关键词后停止提取。

3)将词语转换为词频矩阵,即向量化。

4)统计矩阵中每个词语的 TF-IDF 权值,完成关键词提取和向量转换。

5)计算困惑度,确定 LDA 最优主题个数,并定义函数并输出每个主题里面的前 15 个关键词,完成主题关键词抽取。

6)可视化分析,将 LDA 主题分析结果以直观的形式表现出来,得到交互式的动态图。

主题数与困惑度的折线图如图 6 所示,每个主题下排名前 15 的关键词见表 3,主题 3 示例如图 7 所示。

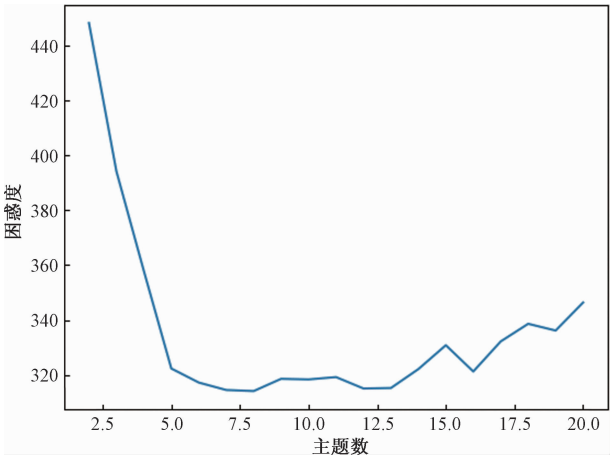
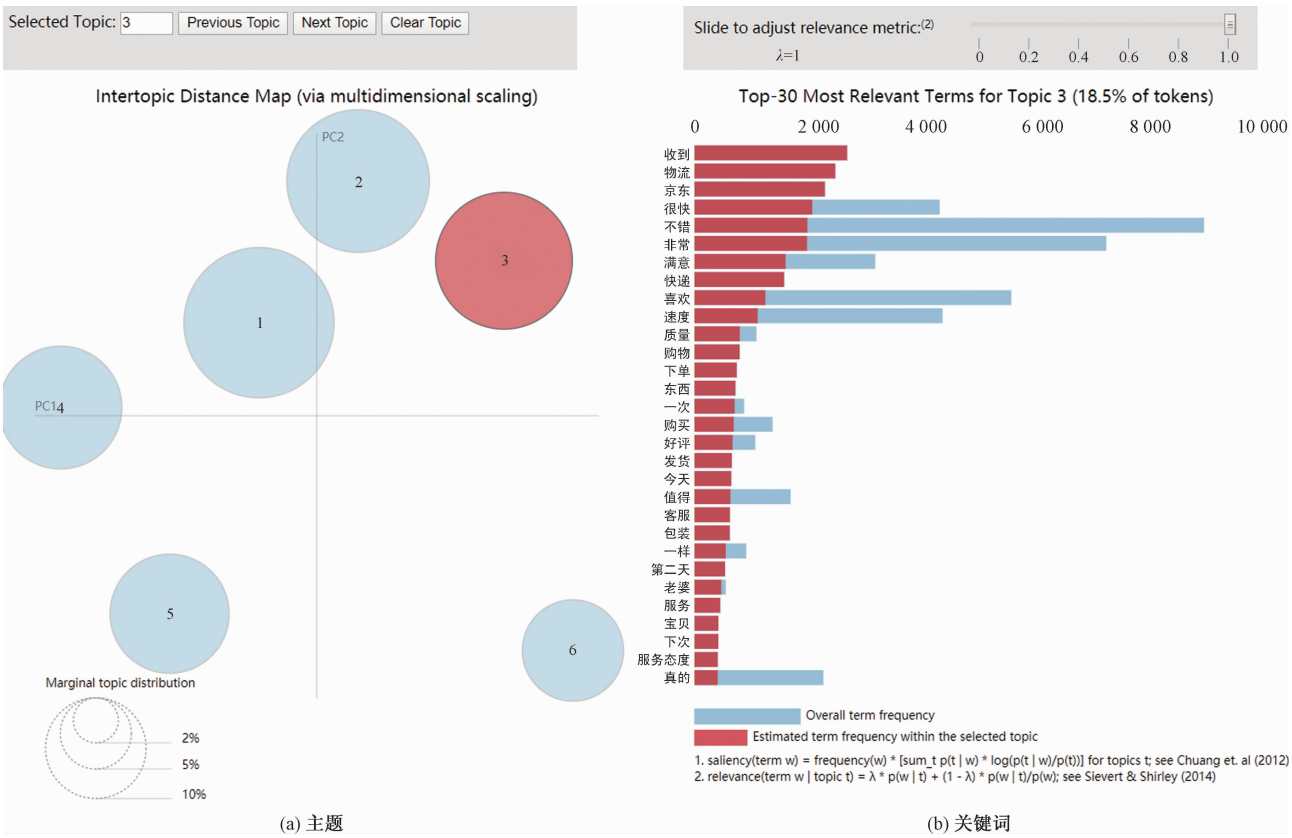


图 6　主题数与困惑度的折线图

表 3　vivo 手机总体评价潜在主题

性价比	电池	物流	拍照/运行	充电/耳机	外观
喜欢	不错	收到	速度	没有	可以
非常	电池	物流	拍照	可以	感觉
不错	可以	京东	运行	就是	不错
满意	流畅	很快	非常	充电	喜欢
vivo	感觉	不错	清晰	问题	外观
值得	性价比	非常	不错	解锁	颜色
性价比	非常	满意	屏幕	耳机	手感
购买	值得	快递	流畅	小时	好用
很高	真的	喜欢	外观	很快	很漂亮
特别	玩游戏	速度	很快	指纹	效果
收到	耐用	质量	效果	感觉	特别
好看	使用	购物	不卡	充满	这款
手感	一天	下单	好看	东西	好用



(a) 主题　　(b) 关键词

图 7　主题 3 结果

图 7(a)表示的是主题，用圆圈代表不同的主题，圆圈的大小代表了每个主题分别包含文章的数量；图 7(b)表示每个主题中常出现的 30 个关键词列表，当鼠标没有悬停在任何主题之上时，这 30 个关键词代表全部文本中提取到的 30 个最重要关键词。当把鼠标悬停在 3，右侧的关键词列表会立即发生变化，红色展示了每个关键词在当前主题下的频率。若模型拟合较好，则图中的圆圈之间将不会重叠，本文从图中看到 6 个主题不存在重叠现象，主题模型拟合的效果较好。

结合表 3 和主题词可视化图进行分析，具体来看每个主题下的信息，主题 1 主要反映手机性价比高；主题 2 主要反映手机的电池耐用；主题 3 主要表现用户对京东的物流速度比较满意；主题 4 表现手机拍照效果、运行速度、屏幕方面的信息；主题 5 主要反映充电、指纹解锁以及耳机问题；主题 6 中外观、颜色主要表现用户对手机外观属性的评价信息，可以、不错、喜欢等词说明用户对手机外观评价

较为积极。综合结果来看,6 个主题无重叠,拟合较好,但该方法对于负面评价主题没有涉及。

3.2.2 基于 Word2vec 和 SVM、LDA 的混合主题分析

将所有评论文本分割成 47 726 条分句,随机选择 1 万条数据,5 人同时对数据人工标注情感极性,积极用"1"表示,消极用"-1"表示,采用少数服从多数的思想确定数据最终的情感极性。

按照训练集与测试集 7∶3 的比例,采用 Word2vec 连续词袋模型对训练集数据构建词向量(每个词用 100 维的向量表示,将句子的词向量平均之后作为该句子的向量);然后对分词之后的数据训练 Word2vec 词向量模型;之后对原有评论数据使用训练好的词向量模型,利用 SVM 训练分类模型,并选择线性核函数将向量映射到空间,判断句子向量映射在哪个超空间里面,即积极还是消极;最后利用训练好的 SVM 分类模型进行情感预测,并对测试集数据预测评估模型效果。评论数据最终被分为正面评价和负面评价文本,再分别进行 LDA 主题分析。正面评论文本被聚成 6 个主题,负面评论被聚成 3 个主题,每个主题下生成 10 个最有可能出现的词语及相应的概率,正面评价潜在主题见表 4,负面评价潜在主题见表 5。

基于 SVM、LDA 的主题分析,选择线性核函数,计算效率较高。在评价分类器效果时,引入了信息检索中的混淆矩阵,进而得到了 SVM 情感分析报告,见表 6。其中分类指标精度和召回率指标考量了分类器对于两个类别的总体的分类效果,由此结合精度和召回率得到了 $F_1=0.963\ 5$,故基于 SVM、LDA 的主题分析结果较好。

正面情感数据集 LDA 主题分析:主题 1 到主题 6 分别主要反映的是手机外观好看、京东物流速度快、拍照清晰、手机充电速度快和电池耐用、手机游戏体验好及性能好、手机性价比高和运行流畅。

负面情感数据集 LDA 主题分析:主题 1 主要反映京东平台客服服务问题,以及手机屏幕存在的一些问题;主题 2 主要反映的是手机屏幕指纹解锁慢的问题,主题 3 反映的是 vivo 手机充电电池发热等问题,以及在京东销售客服上的一些问题。

表 4 vivo 手机正面评价潜在主题

主题	高频关键词									
外观	外观	屏幕	好看	很好	手机	漂亮	颜色	系统	喜欢	不错
物流	喜欢	满意	京东	收到	物流	手机	下单	不错	很快	速度
拍照	手机	特别	拍照	清晰	屏幕	质量	像素	效果	喜欢	好评
充电/快递	电池	值得	很快	快递	充电	购买	耐用	速度	电量	推荐
游戏体验/性能	不错	感觉	游戏	流畅	体验	支持	好用	整体	性能	系统
性价比/运行	性价比	真的	运行	速度	价格	手机	很高	很棒	流畅	超快

表 5 vivo 手机负面评价潜在主题

主题	高频关键词									
客服问题	客服	一个	几天	问题	屏幕	差评	一点	垃圾	京东	手机
屏幕问题	有点	感觉	屏幕	指纹	解锁	慢	失灵	知道	问题	下单
充电发热	充电	有点	电池	评价	问题	发热	感觉	真的	客服	一直

表 6 SVM 情感分析报告

指标	评估值
正确率	0.931 5
精确率	0.976 7
召回率	0.950 5
F1	0.963 5

将基于关键词、基于 LDA 的主题分析与 Word2vec 和 SVM、LDA 混合算法的主题分析结果进行对比分析可以看出:基于关键词的主题分析较为抽象,需要分析人员具备一定的业务知识;基于 LDA 的主题分析相对主题明确、清晰,共得到 6 个互不重叠的主题,主题划分效果较好。基于 Word2vec 和 SVM、LDA 混合算法的主题分析得到两大类主题,每类主题下又细分了子主题。其中正面主题下的 5 个子主题与 LDA 完全相同,负面主题下又细分了 3 个子主题,比 LDA 更加详细、具体,尤其负面主题的分析,对商家的指导意义更为重要。

3.3 主题分析与商家建议

对主题及其中的高频特征词分析可以得出结论,vivo 手机的优势有外观好看、物流速度快、拍照效果好、充电速度快、电池耐用、游戏体验好、性能好、性价比高、运行流畅。用户对 vivo 手机不满意的地方在于京东客服服务态度、手机屏幕指纹解锁

慢、充电电池发热、没有赠送耳机等。

基于京东平台上 vivo 手机的用户评论的 LDA 主题模型分析结果，提出以下建议：①在保持 vivo 手机运行流畅、速度快等优势的基础上，对 vivo 手机在屏幕指纹识别、电池充电上进行改进，从整体上提升 vivo 手机的质量；②加强客服人员的整体素质，提高服务质量，让其在手机行业凸显优势。如果商品本身及服务能够满足以上要求，并辅以恰当的运营手段，在推广手机品牌时才容易和热销的竞品进行竞争。

4 结语

本文基于手机评论大数据，进行可视化分析和主题模型分析，进而挖掘用户评论的焦点和潜在主题信息，并将 LDA 与基于 Word2vec 的 SVM 算法结合进行正、负面主题情感分析。结果表明该方法对用户评论数据的挖掘结果比基于关键词的可视化分析、基于整体数据集的 LDA 主题分析更清晰，能够快速获得用户各方面的反馈，找到手机以及销售平台的具体改进方向，并结合观点挖掘找到用户的不满点，进而确定改进策略。

参考文献

[1] 陈晓美，关心惠. 网络舆情观点提取的 LDA 主题模型方法[J]. 图书情报工作，2015，59(21)：21-26.

[2] 万晓霞. 基于 LDA 模型与聚类的网络新闻热带你话题发现研究[D]. 成都：西华大学，2014.

[3] HU Y，JOHN A，SELIGMANN D D. Event analytics via social media[C]//ACMWorkshop on Social and Behavioural Networked Media Access. New York：ACM，2011：39-44.

[4] 田贤忠，姚明超，顾思义. 基于 BBS-LDA 的论坛主题挖掘[J]. 浙江工业大学学报，2020，48(1)：55-62.

[5] 曾寰，龙小建，刘华. 基于 LDA 及语义相似度的商品评论情感分类研究[J]. 井冈山大学学报，2019，40(4)：46-51.

[6] 谢宗彦，黎嶷，周纯洁. 基于 word2vec 的酒店评论情感分类研究[J]. 北京联合大学学报，2018，32(4)：34-39.

[7] 吴龙峰. 基于 Word2vec 和 LDA 的卷积神经网络文本分类模型[J]. 电脑知识与技术，2019，15(22)：1099-3044.

[8] ZHANG D，XU H，SU Z，et al. Chinese comments sentiment classification based on word2vec and SVM perf[J]. Expert Systems with Application，2015，42(4)：1857-1863.

[9] 张冬雯，杨鹏飞，许云峰. 基于 word2vec 和 SVMperf 的中文评论情感分类研究[J]. 计算机科学，2016，43(S1)：418-421，447.

[10] 王勤勤，张玉红，李培培，等. 基于 word2vec 的跨领域情感分类方法[J]. 计算机应用研究，2018，35(10)：2924-2927.

[11] 杨小平，张中夏，王良，等. 基于 Word2vec 的情感词典自动构建与优化[J]. 计算机科学，2017，44(1)：42-47.

[12] SHARMA A K，CHAURASIA S，SRIVASTAVA D K. Sentimental short sentences classification by using CNN deep learning model with fine tuned Word2vec[J]. Procedia Computer Science，2020，167：1139-1147.

User Comments Subject Analysis Based on Sentiment Analysis Hybrid Model：Taking vivo mobile phone as an example

ZHANG Qiannan

(Basic Teaching Department，Huanghe Jiaotong University，Jiaozuo Henan 454950，China)

Abstract：In order to help merchants understand user needs and product issues，and further improve product and service quality，aiming on on Vivo mobile phone review data，the internal connection between user focus and mobile phone characteristics and the trend of user sentiment tendency are explored through word cloud graph analysis，social semantic network analysis，and public opinion time series visualization analysis of comments data based on keywords. Then，the LDA topic feature analysis has been done. And the hybrid algorithm based on Word2vec，SVM and LDA is put forward to mine the potential topics of the user's positive and negative emotional comments，and get the user's reaction to different aspects of the Vivo mobile phone under different emotional tendencies. The analysis results show that the mining results based on the hybrid algorithm are clearer and more accurate than the visual analysis based on keywords and the topic analysis based on LDA，and the suggestions for businesses are more helpful.

Keywords：data mining；word cloud graph；topic analysis；sentiment analysis；Word2vec