文本挖掘：电商评论分析

报告人：熊欣

时间：2020 年 2 月 10 日

目录

1 问题描述

网络购物的蓬勃发展推动了诸多电商平台的崛起，引发激烈竞争。在竞争大背景下，除了提高商品质量，压低商品价格，了解消费者心声对电商平台变得越来越有必要，其中非常重要的方式就是对消费者的文本评论数据进行内在信息的数据挖掘与分析。获得这些信息，能够有效提升企业或商家的市场竞争力。

本实验以京东六大品牌热水器交易评论数据为研究基础，利用文本数据挖掘技术从文本预处理、词频统计、情感分析等几个方面进行分析，并分析各品牌间的差异，最后给商家提出建议，为电子商务的后期研究打下基础。本实验分析针对京东平台的热水器评论，主要包括以下三个问题：

（1）对每款热水器品牌进行用户情感倾向分析。

（2）从每款热水器品牌的评论文本中挖掘出该款产品的优点和不足。

（3）提炼不同品牌热水器的卖点，通过横向对比，针对目标品牌企业的营销和产品优化提出建议。

2 数据描述

原始数据集：本次实验数据集 eview_data.csv 是包含 AO、格兰仕、海尔、美的、万和、万家乐这六个热水器品牌的京东交易平台的评论数据，数据存储格式为 csv。该原始数据总共有 215032 条数据记录，抓取的京东评论数据集存储结构主要包括 ID、评论、时间、型号、PageUrl 等。其中本实验主要关注于评论文本数据的情感分类和主题分析，故**本实验按照六大热水器品牌进行筛选，提炼抽取出每个热水器品牌的评论文本语料**，数据存储格式为 txt。

停用词典：本次实验应用老师提供的停用词典表 stoplist.txt，这些停用词是为节省存储空间和提高搜索效率，在处理自然语言数据或文本之前或之后会自动过滤掉某些字或词。

用户自定义词典：本次实验应用老师提供的用户自定义词典 Dict.txt，增强文本预处理时的分词准确度，为后续情感分析做好前期基础准备。

大连理工情感词典：在中文领域，大连理工大学信息检索实验室中文情感词汇本体 (http://ir.dlut.edu.cn/) 参考 Ekman 情感模型将情感分为乐 (joy)、惧

(fear)、惊 (surprised)、哀 (sadness)、恶 (disgusted)、怒 (anger)和好 (goodness)7 个大类 21 个小类，在情感分析领域应用广泛。其宗旨是在情感计算领域，为中文文本情感分析和倾向性分析提供一个便捷可靠的辅助手段。本实验应用大连理工情感词典，将乐 (joy)、惊 (surprised)、好 (goodness)这三类情感划分为正向情感类别，将惧 (fear)、哀 (sadness)、恶 (disgusted)、怒 (anger)这四类情感划分为负向情感类别，同时按照正负向情感类别对正向情感词赋值+（情感强度），对负向情感词赋值-（情感强度），构建出实验关键的情感词典。

3 方法描述

本实验目标是通过构建情感词典对每款热水器品牌进行用户正负情感倾向分类，同时利用 LDA 主题模型从每款热水器品牌的评论文本中挖掘出该款产品的优点和不足，最终提炼不同品牌热水器的卖点，通过六大品牌横向对比，针对目标品牌企业的营销和产品优化提出建议。整个实验的方法流程（即老师提供任务流程图）如下图 1 所示。

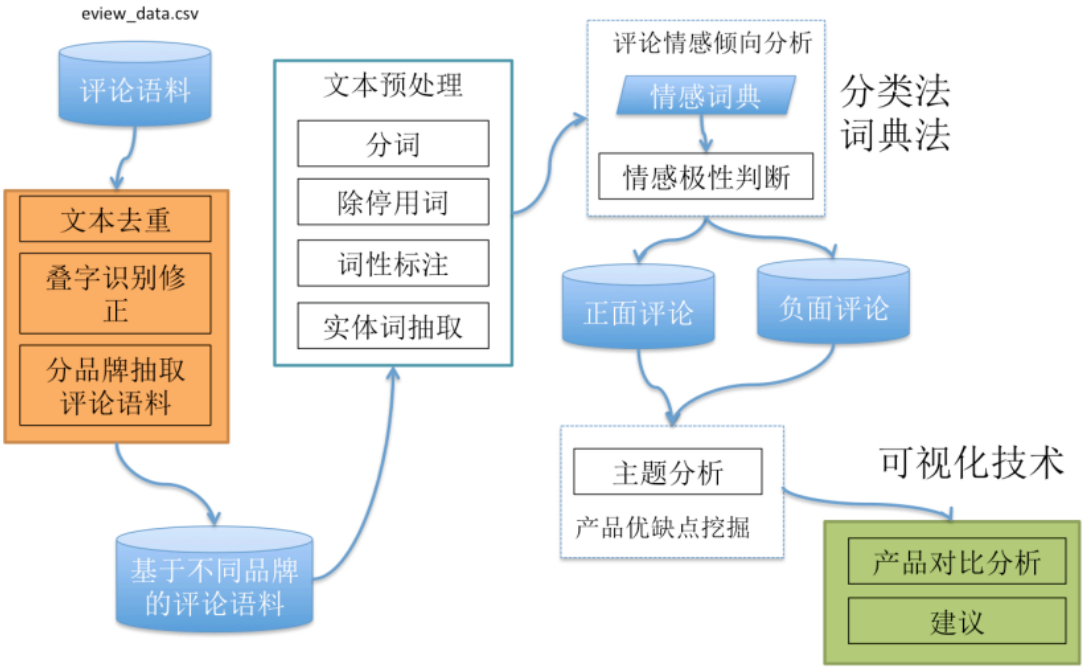

图 1 热水器电商平台评论文本挖掘流程

热水器电商平台评论文本挖掘实验过程主要分为三个部分，第一个部分是对原始评论语料 eview_data.csv 的前期处理，包括文本去重、叠字识别修正、

分品牌抽取生成六大热水器品牌的评论语料。同时针对抽取后的评论文本进行分词、去停用词、词性标注及实体词抽取的预处理操作。

第二个部分是基于构建的情感词典的评论情感倾向分析，此处应用大连理工情感词典，构建出实验需要的情感词典。并基于情感词典对各品牌的评论文本进行情感极性判断，划分正面评论与负面评论两类。

第三个部分是基于主题模型的产品优缺点挖掘，应用 LDA 模型对六大热水器品牌的正负面评论文本进行主题分析，总结各品牌热水器的优点和不足，并结合相关市场信息可视化呈现六大热水器品牌的产品对比结果，并针对分析结果给出相应的建议对策。

4 数据分析与建模

本章节主要对原始数据进行探索性分析，包括文本去重、叠字识别修正等数据清洗操作，分词、除停用词、词性标注、实体词抽取等文本预处理操作，然后基于大连理工情感词典对各品牌的评论语料进行情感极性判断，划分为正面评论与负面评论，并分别对其进行 LDA 主题分析，挖掘出各品牌热水器产品的优缺点。

4.1 数据清洗

4.1.1 文本去重

文本去重，就是去除文本评论数据中重复的部分。本实验需要进行文本去重的原因有：京东电商平台为了避免一些客户长时间不评论，系统会自动给超过规定时间仍未作出评论的用户做出好评，从而出现大量重复评论；同一个用户可能因为购买多款热水器而出现重复的评论等等。

目前已有很多文本去重算法，大多都是先通过计算文本之间的相似度，再以此为基础进行去重，包括编辑距离去重，simhash 算法去重等，这些会使得我们去除一些相近的表达，造成错删存在缺陷。这一类相对复杂的文本去重算法容易去除，则本实验考虑一些相对简单的文本去重思路，采用比较删除法，两两比对完全重复的语料直接删除，尽量保留有用的评论。由于前期已经将原始 eview_data 评论语料库按照热水器六大品牌提炼抽取出每个品牌的评论文本，故依次对这六个热水器品牌评论文本语料进行文本去重，其中 AO 史密斯热水

器评论语料经过文本去重后剩余 9357 条评论，格兰仕热水器评论语料经过文本去重后剩余 26049 条评论，海尔热水器评论语料经过文本去重后剩余 66505 条评论，美的热水器评论语料经过文本去重后剩余 53048 条评论，万和热水器评论语料经过文本去重后剩余 17881 条评论，万家乐热水器评论语料经过文本去重后剩余 6579 条评论。

4.1.2 叠词识别修正

由于热水器电商品牌的文本评论数据质量参差不齐，没有意义的文本数据较多，通过文本去重可以删除很多没有意义的评论文本还远远不够，还需要对去重后的评论文本进行叠词识别修正。例如“非常好非常好非常好非常好非常好”转化为“非常好”、“哈哈哈哈哈哈哈哈哈”转化为“哈”、“可以可以可以可以可以”转化为“可以”等。根据叠词识别修正规则，可以完成对开头连续重复的处理，类似的也可以对处理过的文本再进行以此结尾连续重复的叠词识别修正，机械式压缩去词，算法思想是相近的，只是从尾部开始读词。从结尾开始的处理结束后就得到修正压缩去词完成的精简评论文本语料。

4.1.3 短句删除

由于字数越少所能够表达的意思越少，要想表达一些相关的意思一定要有相应量的字数，过少字数的评论必然是没有任何有意义的评论。为此，就要删除掉过短的评论文本数据，以去除掉没有意义的评论。显然短句删除最重要的环节就是保留的评论字数下限的确定，没有精确的标准，可以结合特定语料来确定，一般 4-8 个国际字符都是较为合理的下限。本实验中设置经过数据清洗后得到的评论语料若小于等于 4 个国际字符，则将该语料删去。经过数据清洗后保留下的语料文本如下图 2 所示。

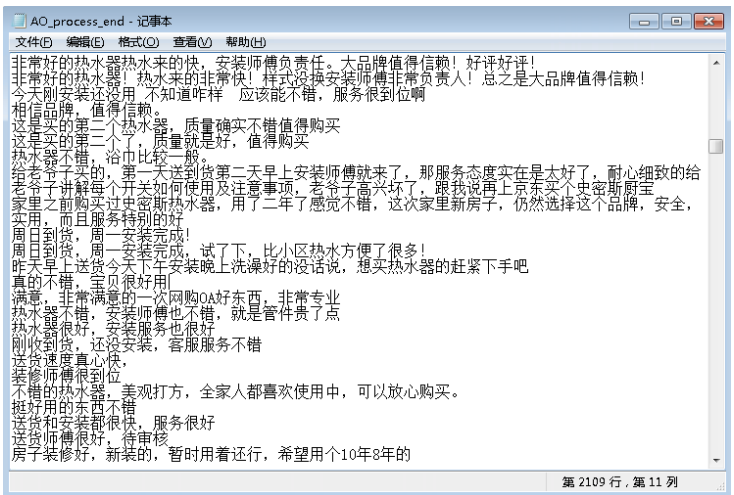

图 2 数据清洗后的 AO 热水器评论语料

4.2 文本预处理

4.2.1 分词及除停用词

由于中文只有字、句和段落能够通过明显的分界符进行简单的划界，而对于词和词组来说，它没有一个形式上的分界符。因此，进行中文文本挖掘时，首先应对文本分词，即将连续的字序列按照一定的规范重新组合成词序列的过程。分词结果的准确性对后续文本挖掘算法有着不可忽视的影响，如果分词效果不佳，即使后续算法优秀也无法实现理想的效果。本实验采用 Python 的中文分词包 jieba（结巴分词），对六大热水器品牌的商品评论文本数据进行中文分词。同时对分词后的评论语料进行除停用词操作，主要包括英文字符、数字、数学字符、标点符号及使用频率特高的单汉字等的去除，能够节省存储空间和提高搜索效率。

4.2.2 词性标注及实体词抽取

六个品牌的评论文本语料经过分词、去停用词等预处理后，利用 jieba.analyse.extract_tags 进行词性标注及实体词抽取，**保留名词、形容词、动词及副词作为候选**。该步骤是为了提高后续词典匹配的精确度，对候选词项进行情感词辨别，依据词典标注赋予词条情感类标及情感强度。经过文本预处理后保留下的语料文本如下图 3 所示。

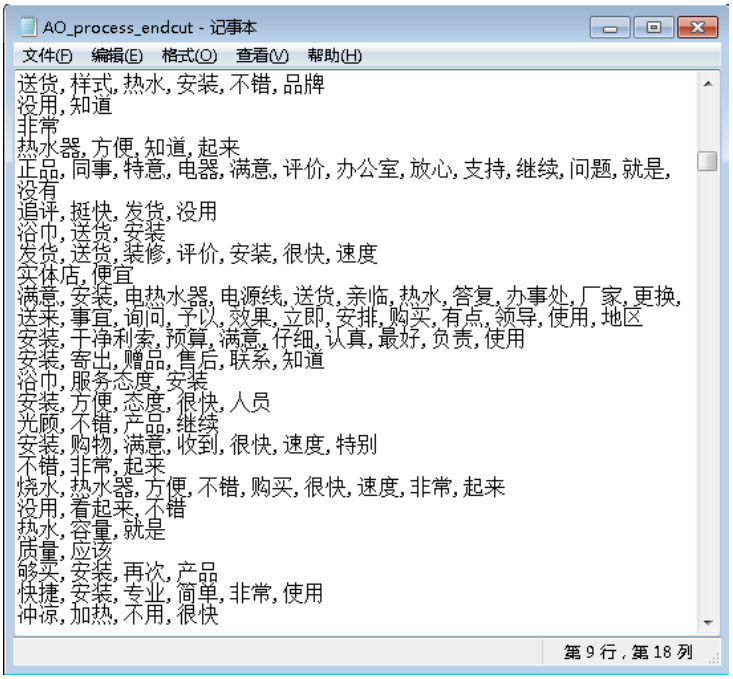

图 3 数据清洗后的 AO 热水器评论语料

4.3 基于情感词典的情感分类

传统的基于情感词典的文本情感分类，是对人的记忆和判断思维的最简单的模拟。首先通过学习来记忆一些基本词汇，如否定词语有“不”，积极词语有“喜欢”、“爱”，消极词语有“讨厌”、“恨”等，从而在大脑中形成一个基本的语料库。然后，我们再对输入的句子进行最直接的拆分，看看我们所记忆的词汇表中是否存在相应的词语，然后根据这个词语的类别来判断情感，比如“我喜欢数学”，“喜欢”这个词在我们所记忆的积极词汇表中，所以我们判断它具有积极的情感。在本小节中，需要对各个品牌热水器的评论文本进行情感极性判断，即通过评论情感倾向分析来将各品牌评论文本划分为正面评论和负面评论。

本实验**应用大连理工情感词典，将乐 (joy)、惊 (surprised)、好 (goodness)这三类情感划分为正向情感类别，将惧 (fear)、哀 (sadness)、恶 (disgusted)、怒 (anger)这四类情感划分为负向情感类别，同时按照正负向情感类别对正向情感词赋值+（情感强度），对负向情感词赋值-（情感强度），构建出实验需要的情感词典**。由于前期已经对热水器品牌评论文本语料进行了数据清洗和预处理得到较工整的文本语料集，故本文思路是直接比对预处理后的评论文本候选词和构建的情感词典集。若评论文本中的候选词与情感词典中的词语匹配，则给该句评论文本赋予对应的情感值；若一条评论文本包含多个情感词汇，则采取简单线性叠加原理，计算得出该条评论文本的最终情感值。最后，根据计算得到的总情感值的正负性来判断句子的情感取向。可以得到如 AO 热水器的正面评论和负面评论文本如下图 4 和 5 所示。

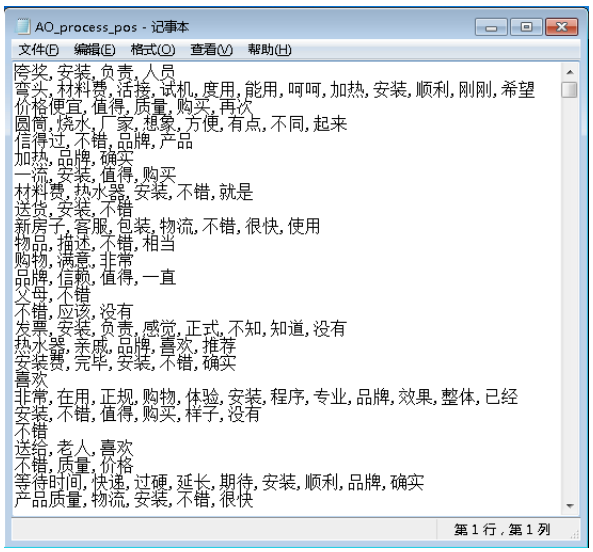

图 4 AO 热水器的正面评论文本

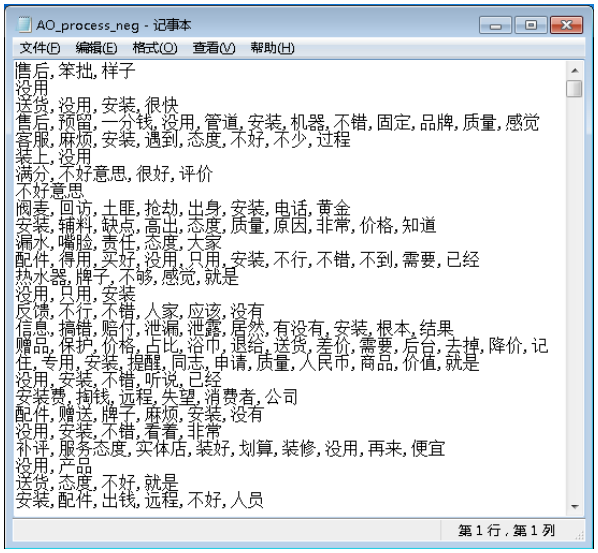

图 5 AO 热水器的负面评论文本

4.4 基于 LDA 主题分析

在对评论文本进行情感词识别之后，需要对正负面情感评论内容描述进行主题聚类，以更好地凝聚评论内容中蕴含的情感特征。LDA 模型是一种主题模型，它可以将文档集中的每篇文档的主题以概率分布的形式给出，从而通过分析一些文档抽取出它们的主题（分布）出来后，便可以根据主题（分布）进行主题聚类或文本分类。同时，它是一种典型的词袋模型，即一篇文档是由一组词构成，词与词之间没有先后顺序的关系。本实验采用隐主题模型 LDA 对正负面评论文本进行主题聚类，以挖掘六个热水器品牌评论中的更多信息。

4.4.1 情感主题建模

首先建立情感主题模型需要进行模型的优选参数实验。依据前人的实验经验，LDA 的超参数 α 一般设置为 50/K（K 为 topic 的数量），β 一般 取 0.01 时模型效果较好，收敛较快。故本研究需要确定最佳主题数 K 值，从而确定超参数 α 和 β 的值。目前使用较多的确定主题数 K 值的方法是通过计算不同主题数对应的 perplexity 值，其中 perplexity 是一种信息理论的测量方法，b 的 perplexity 值定义为基于 b 的熵的能量（b 可以是一个概率分布，或者概率模型），通常用于概率模型的比较，该方法适用于长文本主题模型构建。考虑到本实验预处理后的评论文本属于短文本，故本实验**利用 LDAvis 通过动态参数调整，来确定最佳主题数 K 值**。pyLDAvis 模块是 python 中一个对 LDA 主题模型算法的可视化模块，可以通过可视化图形式来直观查看 LDA 模型的聚类效果。

以 AO 热水器品牌的正面评论文本为例，如图 6 所示，**当 K=3 时，构建的 LDA 模型中的各个主题之间相离较远，主题凝聚效果较好**。

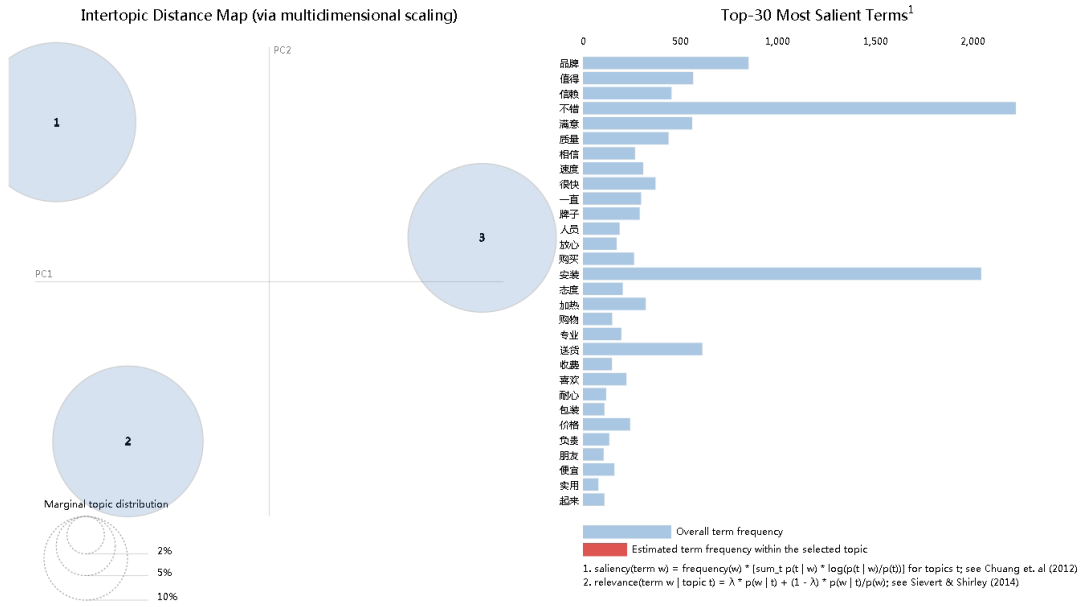

图 6 AO 热水器的正面评论文本

4.4.2 基于 LDA 的评论情感主题

根据上文情感主题建模结果，通过 pyLDAvis 可视化模块选择 K=3 时构建评论文本主题聚类模型。按照六个热水器品牌，**对其正负面评论分别聚类划分为 3 个潜在主题，每个主题下提取 10 个高频特征词，从而挖掘出潜在主题下的深层含义**。

（1）AO 热水器评论情感主题

经过 LDA 主题分析，如下表 1 和 2 显示了 AO 热水器正面和负面评论文本中的潜在主题。

表 1 AO 正面评论潜在主题高频特征词

主题 1		**主题 2**		**主题 3**	
安装	送货	满意	很快	不错	热水器
就是	人员	不错	速度	品牌	安装
好评	负责	非常	送货	信赖	一直
热水器	不错	购买	安装	值得	牌子
态度	服务态度	质量	很好	相信	喜欢

表 2 AO 负面评论潜在主题高频特征词

主题 1	**主题 2**	**主题 3**

没用	知道	看着	配件	没用	就是
应该	热水器	没用	便宜	安装	看起来
不错	评价	安装	比较	不错	已经
安装	装修	不错	有点	人员	产品
不好	麻烦	安装费	大家	感觉	装上

根据对 AO 热水器正面评论的三个潜在主题的高频特征词分析，可以发现主题 1 的高频特征词包括送货、人员、服务态度等，主要反映 AO 热水器的客服或送货人员的服务态度较好；主题 2 的高频特征词包括安装、质量、满意等，热门关注点主要是 AO 热水器的质量较好值得购买；主题 3 的高频特征词包括品牌、喜欢、信赖等，主要反映 AO 热水器的品牌值得信赖受到喜欢。

根据对 AO 热水器负面评论的三个潜在主题的高频特征词分析，可以发现主题 1 的高频特征词包括安装、装修、麻烦等，主要反映 AO 热水器的安装较麻烦；主题 2 的高频特征词包括配件、有点等，热门关注点主要是 AO 热水器的配件可能存在不足；主题 3 的高频特征词包括人员、产品等，主要反映 AO 热水器产品自身可能存在些问题。

（2）格兰仕热水器评论情感主题

经过 LDA 主题分析，如下表 3 和 4 显示了格兰仕热水器正面和负面评论文本中的潜在主题。

表 3 格兰仕正面评论潜在主题高频特征词

主题 1		**主题 2**		**主题 3**	
不错	便宜	送货	起来	安装	朋友
满意	安装	不错	加热	就是	热水器
价格	很好	很快	方便	性价比	没有
非常	实用	速度	安装	值得	配件
感觉	质量	喜欢	效果	购买	品牌

表 4 格兰仕负面评论潜在主题高频特征词

主题 1		**主题 2**		**主题 3**	
没用	知道	麻烦	有点	安装	不行
不错	评价	就是	不错	售后	态度
应该	还行	不好	比较	没有	差劲
感觉	不合理	不够	速度	热水器	客服

安装	凑合	安装	送货	电话	人员

根据对格兰仕热水器正面评论的三个潜在主题的高频特征词分析，可以发现主题 1 的高频特征词包括质量、实用、满意等，主要反映格兰仕热水器的质量较好；主题 2 的高频特征词包括安装、方便、很快、加热等，热门关注点主要是格兰仕热水器的加热速度快且安装方便；主题 3 的高频特征词包括性价比、品牌、值得等，主要反映格兰仕热水器品牌值得信赖性价比高。

根据对格兰仕热水器负面评论的三个潜在主题的高频特征词分析，可以发现主题 1 的高频特征词包括不合理、凑合等，主要反映格兰仕热水器使用体验一般；主题 2 的高频特征词包括麻烦、安装、不好等，热门关注点主要是格兰仕热水器的安装麻烦不好；主题 3 的高频特征词包括售后、客服、态度、差劲等，主要反映格兰仕热水器的售后和京东客服人员的态度较差。

（3）海尔热水器评论情感主题

经过 LDA 主题分析，如下表 5 和 6 显示了海尔热水器正面和负面评论文本中的潜在主题。

表 5 海尔正面评论潜在主题高频特征词

主题 1		**主题 2**		**主题 3**	
不错	送货	就是	有点	不错	品牌
很快	性价比	安装	收费	满意	信赖
加热	外观	起来	没有	安装	非常
速度	便宜	安装费	材料费	值得	价格
喜欢	支持	朋友	推荐	质量	感觉

表 6 海尔负面评论潜在主题高频特征词

主题 1		**主题 2**		**主题 3**	
没用	应该	安装	没有	不够	热水
不错	知道	就是	送货	加热	方便
发现	不好	麻烦	材料费	有点	速度
缺点	暂时	售后	安装费	就是	不好
安装	送货	收费	失望	保温	凑合

根据对海尔热水器正面评论的三个潜在主题的高频特征词分析，可以发现主题 1 的高频特征词包括性价比、外观、加热等，主要反映海尔热水器的性价比较高、外观好；主题 2 的高频特征词包括安装、材料费、推荐等，热门关

注点主要是海尔热水器的安装材料费不错；主题 3 的高频特征词包括品牌、质量、信赖等，主要反映海尔热水器的品牌质量值得信赖。

根据对海尔热水器负面评论的三个潜在主题的高频特征词分析，可以发现主题 1 的高频特征词包括安装、送货、缺点等，主要反映海尔热水器的安装送货存在不足；主题 2 的高频特征词包括售后、材料费、安装费等，热门关注点主要是海尔热水器的安装售后费用服务仍不够好；主题 3 的高频特征词包括保温、加热、速度、不好等，主要反映海尔热水器的加热保温效果不佳。

（4）美的热水器评论情感主题

经过 LDA 主题分析，如下表 7 和 8 显示了美的热水器正面和负面评论文本中的潜在主题。

表 7 美的正面评论潜在主题高频特征词

主题 1		**主题 2**		**主题 3**	
不错	热水器	满意	品牌	不错	起来
安装	性价比	不错	送货	很快	方便
就是	便宜	值得	信赖	速度	安装
感觉	价格	质量	购买	加热	朋友
喜欢	安装费	非常	很好	使用	外观

表 8 美的负面评论潜在主题高频特征词

主题 1		**主题 2**		**主题 3**	
没用	加热	安装	没有	缺点	看着
不错	麻烦	热水器	态度	发现	可惜
知道	应该	售后	人员	不好	安装
不够	有点	就是	安装费	暂时	不足
安装	就是	不行	材料	没有	不错

根据对美的热水器正面评论的三个潜在主题的高频特征词分析，可以发现主题 1 的高频特征词包括价格、便宜、性价比、安装费等，主要反映美的热水器的性价比较高，价格便宜适合大众购买；主题 2 的高频特征词包括品牌、满意、质量、值得、购买等，主要反映美的热水器的品牌还是很受大众认可和满意，质量较好值得购买；主题 3 的高频特征词包括加热、速度、使用、安装、方便等，即热门关注点主要是美的热水器的加热速度快，安装和使用较为方便。

根据对美的热水器负面评论的三个潜在主题的高频特征词分析，可以发现主题 1 的高频特征词包括安装、不够、麻烦等，主要反映美的热水器的安装过程较为麻烦；主题 2 的高频特征词包括售后、态度、人员等，热门关注点主要是美的热水器的京东售后服务态度不好；主题 3 的高频特征词包括缺点、发现、可惜等，主要反映美的热水器自身可能存在不满足用户需求的缺点。

（5）万和热水器评论情感主题

经过 LDA 主题分析，如下表 9 和 10 显示了万和热水器正面和负面评论文本中的潜在主题。

表 9 万和正面评论潜在主题高频特征词

主题 1		**主题 2**		**主题 3**	
安装	安装费	不错	方便	满意	很快
就是	热水器	质量	便宜	送货	很好
值得	品牌	性价比	加热	不错	速度
感觉	购买	使用	价格	朋友	一直
没有	配件	起来	喜欢	非常	产品

表 10 万和负面评论潜在主题高频特征词

主题 1		**主题 2**		**主题 3**	
不好	麻烦	没用	暂时	安装	就是
安装	不行	不错	应该	没有	投诉
就是	有点	发现	使用	配件	售后
安装费	售后	缺点	没有	热水器	购买
热水器	比较	知道	感觉	安装费	差劲

根据对万和热水器正面评论的三个潜在主题的高频特征词分析，可以发现主题 1 的高频特征词包括值得、配件、购买等，主要反映万和热水器配件齐全值得购买；主题 2 的高频特征词包括性价比、便宜、价格、喜欢等，热门关注点主要是万和热水器的价格便宜性价比较高；主题 3 的高频特征词包括送货、满意、很快等，主要反映万和热水器的送货很快产品满意。

根据对万和热水器负面评论的三个潜在主题的高频特征词分析，可以发现主题 1 的高频特征词包括安装、不行、安装费等，主要反映万和热水器的安装麻烦且安装费用较高；主题 2 的高频特征词包括发现、缺点、使用等，主要

反映了万和热水器自身可能存在些许问题或缺点；主题 3 的高频特征词包括差劲、投诉、售后等，热门关注点是万和热水器的售后服务较差遭到投诉。

（6）万家乐热水器评论情感主题

经过 LDA 主题分析，如下表 11 和 12 显示了万家乐热水器正面和负面评论文本中的潜在主题。

表 11 万家乐正面评论潜在主题高频特征词

主题 1		**主题 2**		**主题 3**	
热水器	安装	送货	实用	不错	朋友
就是	品牌	安装	不错	质量	加热
不错	起来	很快	速度	满意	喜欢
值得	感觉	产品	应该	性价比	外观
价格	便宜	没有	问题	非常	保温

表 12 万家乐负面评论潜在主题高频特征词

主题 1		**主题 2**		**主题 3**	
缺点	暂时	没用	价格	不够	送货
发现	不行	不好	质量	安装	收费
安装	购买	不错	有点	麻烦	热水器
没有	投诉	知道	水阀	就是	比较
热水器	使用	安装	便宜	应该	个人

根据对万家乐热水器正面评论的三个潜在主题的高频特征词分析，可以发现主题 1 的高频特征词包括价格、值得、便宜、品牌等，主要反映万家乐热水器品牌受到认可且价格便宜合适；主题 2 的高频特征词包括送货、安装、实用、速度等，热门关注点主要是万家乐热水器的安装送货速度较快；主题 3 的高频特征词包括质量、满意、朋友、喜欢等，主要反映万家乐热水器质量较好受到很多人喜欢。

根据对万家乐热水器负面评论的三个潜在主题的高频特征词分析，可以发现主题 1 的高频特征词包括投诉、缺点、发现等，主要反映用户在使用万家乐热水器时出现问题申请投诉现象；主题 2 的高频特征词包括水阀、安装、质量等，热门关注点主要是万家乐热水器的硬件质量方面如水阀等可能存在问题；主题 3 的高频特征词包括送货、收费、个人、麻烦等，主要反映万家乐热水器的送货或安装收费可能较高。

5 实验结果讨论

本章节主要是从宏观整体角度对六大热水器品牌进行产品对比分析，并根据分析讨论结果提出针对性建议对策。

5.1 产品对比分析

通过分析六大热水器品牌的京东平台用户评论情感分布与主题，能够发现各品牌的评论文本数量如下图 7 所示。从中可以得出，**美的和海尔这两大畅销品牌是最受欢迎的，销量占据总销量的百分之六七十**，这与两大品牌的市场价格适中及产品性价比较高有关；格兰仕和万和热水器的销量适中，占据一定市场份额；而 AO 史密斯和万家乐热水器销量较少，这可能与 AO 史密斯热水器价位较高主打中高端市场有关，万家乐热水器的品牌认可度仍不及其他几个品牌，导致市场销量较低。同时各品牌间的正负面情感分布比例差异如下图 8 所示，即正面情感所占比例从大到小排列为：**AO 史密斯>海尔>美的>格兰仕>万家乐>万和**，即 AO 史密斯、海尔以及美的这三种品牌的热水器各方面质量服务等更受用户的喜爱，格兰仕、万家乐及万和这三种品牌的热水器表现较前三种口碑较差一些。这说明，虽然 AO 史密斯热水器的价格稍高，但其质量保障、产品性能和售后服务等各方面表现优异，仍然能够得到较好的用户情感反馈。而万家乐、万和热水器在产品价格上较实惠，但可能存在产品质量或售后服务等方面问题影响用户的使用体验。

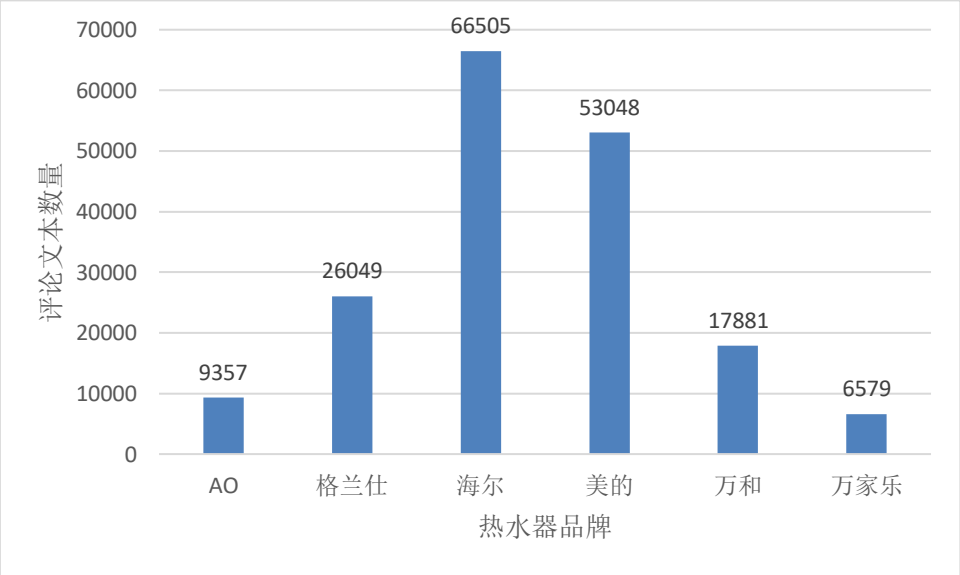

图 7 六大品牌热水器评论文本数量

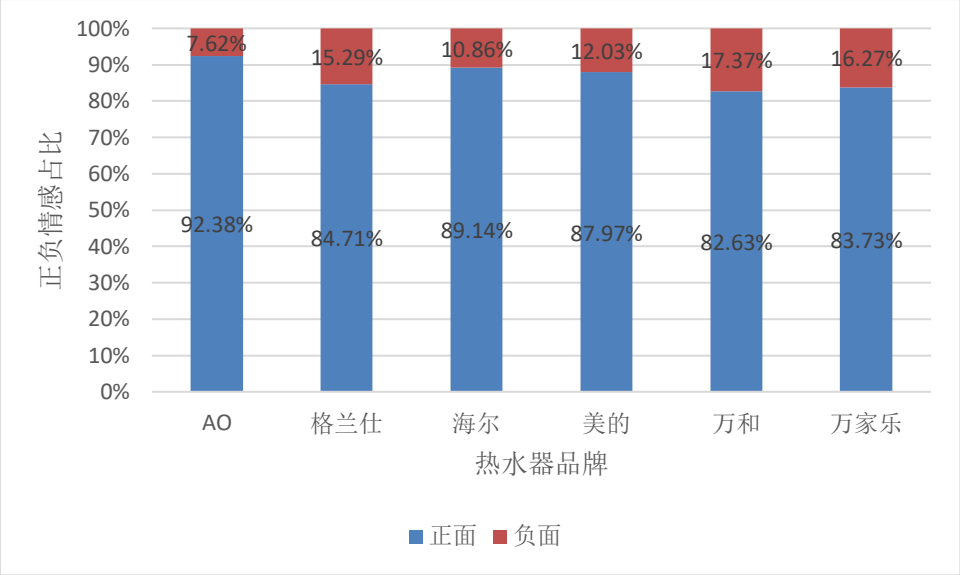

图 8 六大热水器品牌正负面评论情感倾向对比

综合以上对六大热水器品牌的情感主题及其中高频特征词，对京东这六大品牌热水器的用户正面和负面的评论分析结果表明，用户对于热水器反馈的重点主要包括价格、性价比、加热速度、保温性、安装费用、售后服务等方面。本实验从六大热水器品牌的正负面评论中剔除“热水器”无效高频特征词，绘制出正负面评论的词云图。结合 4 部分的主题建模分析结果，本实验针对用户关注重点问题进行品牌横向对比，如下表 13 所示。**AO 史密斯、格兰仕、海尔、美的热水器的品牌效应较好受到大众信赖，且质量口碑较好；格兰仕、海尔、美的、万和热水器的性价比较高；用户在热水器的安装送货方面普遍存在争议，褒贬不一，体现在安装麻烦程度和安装费用合理性上；AO 史密斯、格兰仕、美的热水器的加热/保温效果较好；关于售后服务态度方面，AO 史密斯品牌表现良好，其他品牌均在该方面表现较差。**

表 13 六大热水器品牌横向对比分析

关注点 品牌	品牌效应	价格/性价比	质量	安装/送货	加热/保温	售后服务态度
AO 史密斯	信赖	价格较高	较好	较麻烦	效果较好	好
格兰仕	信赖	性价比高	较好	一般	效果较好	较差
海尔	信赖	性价比高	较好	费用较高	效果不佳	较差
美的	信赖	性价比高	较好	一般	效果较好	较差
万和	一般	性价比高	一般	一般	效果一般	较差
万家乐	一般	价格便宜	一般	较麻烦	效果不佳	一般

图 9 AO 正面（左）和负面（右）评论潜在主题特征词云

图 10 格兰仕正面（左）和负面（右）评论潜在主题特征词云

图 11 海尔正面（左）和负面（右）评论潜在主题特征词云

图 12 美的正面（左）和负面（右）评论潜在主题特征词云

图 13 万和正面（左）和负面（右）评论潜在主题特征词云

图 14 万家乐正面（左）和负面（右）评论潜在主题特征词云

5.2 建议对策

通过文本挖掘，根据对京东平台上六大热水器品牌的用户评论情况进行情感分类和 LDA 主题模型分析，我们对这六大热水器品牌提出以下建议对策。

（1）针对 AO 史密斯热水器，由于 AO 史密斯品牌热水器主打中高端热水器市场，价位相对较高，为了提升用户口碑，需要更加**关注热水器产品质量**

问题。如增强热水器的安装配送服务质量，让热水器安装更加便捷化；改进热水器加热方式使得其加热速度更快、保温效果更好。

（2）针对格兰仕、海尔和美的热水器，这三种品牌的热水器属于热水器市场中销量最好、用户认可接受度最高的三种，主打性价比较高的优势。在关注热水器产品自身的质量问题及价格实惠等优点的同时，需要**提升安装人员或客服人员的整体服务素质，提高服务质量**。如安装费用收取明文细则，并进行公开透明，减少安装过程中的乱收费问题。适当降低安装费用和材料费用，以此在中坚力量品牌中凸显优势。

（3）针对万家乐和万和热水器，这两种品牌的热水器在热水器市场中销量较少一些，价格也较便宜。如果想在市场中占据更大的份额，仍**需要不断提高品牌在用户心中的知名度，同时也要提高热水器产品自身质量及售后服务质量**，对热水器进行不断改进，从整体上提升热水器的质量。

6 实验总结

本实验对原始评论语料进行文本去重、叠字识别修正、分品牌抽取生成六大热水器品牌评论语料。同时针对抽取后的评论文本进行分词、去停用词、词性标注及实体词抽取的预处理操作。然后应用大连理工情感词典，构建出实验需要的情感词典。并基于情感词典对各品牌的评论文本进行情感极性判断，划分正面评论与负面评论两类。最后应用 LDA 模型对六大热水器品牌的正负面评论文本进行主题分析，挖掘出各品牌热水器的优点和不足，并结合相关市场信息可视化呈现六大热水器品牌的产品对比结果，并针对分析结果给出相应的建议对策。

后续实验可以从以下几个方面进一步深入分析：优化基于情感词典的情感分类方法；优化 LDA 模型的参数选择，改进文本表示方法；深层次分析各品牌热水器的优缺点，针对每一种品牌热水器给出特定的建议对策。

附：程序代码

见文件“电商评论分析-熊欣.py”。