第三届“泰迪杯”

全国大学生数据挖掘竞赛

优

秀

作

品

作品名称：基于电商平台家电设备的消费者评论数据挖掘分析

荣获奖项：二等奖

作品单位：暨南大学

作品成员：邓伟雄　童雪玉　黄国南

指导教师：张元标

基于电商平台家电设备的消费者需求及产品数据挖掘

摘 要

本文通过对电商评论数据的处理和分析，构建了垃圾评论识别模型、基于 RAE 词向量自编码的 SVM 文本情感极性分析模型和产品优劣势分析模型进行文本挖掘，最后基于对淘宝指数和百度指数的提取与分析，构建了用户购买行为的挖掘模型。

针对垃圾评论的识别问题，将垃圾评论归为无关信息、水军评论和系统默认好评三种，并从根据不同的分类特征制定规则予以剔除。

针对评论情感分析，尝试使用半监督的深层学习 RAE 模型，采用 word2vec 工具对 8 万多条评论进行训练得到词向量，再对评论进行情感极性分类，然后从情感的积极方提炼出产品的优势，从情感的消极方提取产品的劣势，但由于其对于不同软件的接口封装较难转移，参数繁多且较难设定和偏置函数无法获得等原因，进而改用基于 RAE 的递归自编码模型的有监督的 SVM 模型，进行情感极性识别，通过手工标示 400 条评论的情感极性，进而训练 SVM 模型，使其对剩下的评论进行情感极性分类，结果显示情感分类的正确率达 85%。

针对产品优劣势分析，由于消极情感只占总评价数的 0.28%，样本过小，因此从消极的情感方提取产品劣势并不可行，转而使用用户关注度分析的方法对产品属性下的用户满意度进行统计分析，通过词频统计提炼出产品的优劣势所在。

针对用户购买行为的挖掘，先确定一组搜索关键字，然后爬取对应关键字下的日搜索量，搜索人群年龄性别及消费能力等分布，进行确定产品的主要消费人群及其消费关注点

关键词：

词向量　递归自编码　SVM 模型　情感极性分析

The data mining based on the electric business platform about consumers' demands and products characters

Abstract:

To deeply mine the comments of ecomercial products, this paper aims to build the model of invalid comments recognition, the SVM text emotional polarity analysis model based on RAE auto coding and then distinguishes the advantages and disadvantages via texts analysis. At last, it grabs and analyzes the Taobao index and Baidu index, building the purchase behavior mining model.

In the invalid comments recognition model, it first labels three kinds of invalid information, like irrelevant comments, posters comments and system comments. Then separate these information by their own characters.

As for the emotional polarity analysis, this paper tried the semi-supervised deep learning RAE model at first, using toolbox word2vec to initial eighty thousands term vectors separated from our comment list. Then classified the comments based on these vectors with RAE, obtaining the advantages from the positive comments and the disadvantages from the negative. However, given the difficulty to transfer packages among different softwares, the numeric unknown parameters and offset function, it tries another supervised approaches SVM model based on RAE auto coding. By handmade labeling four hundreds comments with emotional polarity to train the SVM, then use the well-trained models to classify the rest comments, showing that it has an 85% accuracy.

In the advantages and disadvantages analysis model, the negative comments just account for 0.28%, a small scale, making the plan to obtain negative information infeasible. Hence it's to be transferred into the approaches to analyze the customers' attention to the properties of the product, count the satisfaction degree under each property. Then get the advantages and disadvantages via word frequency statistics.

In the customers' behavior mining part, it's to set a group of keywords, used to get the search clicks under each terms. And then mining the age and consumption level to get the main consumer groups and their focus points.

Key words:

Term vectors Recursive since the coding SVM model Emotional polarity analysis

目　录

1. 挖掘目标

本次建模目标是利用在各大电商平台抓取下来的真实评论数据，首先进行水军和随意发表的评论的识别与剔除，再采用数据挖掘技术，构建基于 RAE 自编码的 SVM 模型，进行有监督的分析，即先手工进行部分评论的情感极性标识作为训练语料，得到用户评论中所包含的情感极性。从而可以在情感极性为正的句子中提取产品优势和用户购买的原因，在情感极性为负的句子中提取产品劣势和个性化需求。从各大电商网站中重新爬取商家推荐的产品优势，再与我们从评论中提取出从各类产品优势中提炼不同产品的差异化卖点。最后，根据百度指数和淘宝指数对关键词热水器和净水机进行查找，能够找到热水器和净水机的消费人群，人群购买的关注点及搜索的关键字。

2. 分析方法与过程

2.1 总体流程

本部分使用一个总体流程图描述建模方法及过程，并对各部分进行简要说明。

流程图见**图 1**.

本用例主要包括如下步骤：

步骤一：使用火车浏览器爬取相关数据，获得初始数据。

步骤二：对评论的可信度进行分析可得评论中包含三类垃圾评论，制定规则分别对三类垃圾评论进行处理。

步骤三：使用 R 语言对热水器和净水机的评论进行切词，将整个句子切成独立的词块。

步骤四：使用 word2vec 将已经切碎的词转化成词向量。

步骤五：构建 SVM 模型，同时进行手工标记样本的情感极性及产品属性。将手工标记的评论数据用于三方面：模型的训练、模型的准确度检测及模型的调整。

步骤六：对模型进行优化重构后输入词向量重新组建的句向量，利用经过训练的 SVM 模型的处理输出情感极性。

步骤七：通过对步骤六的计算可得非好评在用户评论中的比重很小，因此可以通过人工统计的方法找寻产品的优劣势。

步骤八：使用 SVM 模型统计评论中用户对产品性能的认可，进而可以找寻各品牌间产品的差异，构建四分图可以得到产品的优势点。

步骤九：对百度指数和淘宝指数进行分析，得到产品的目标消费人群、用户购买的关注点及用户购买的关注点及主要消费人群。

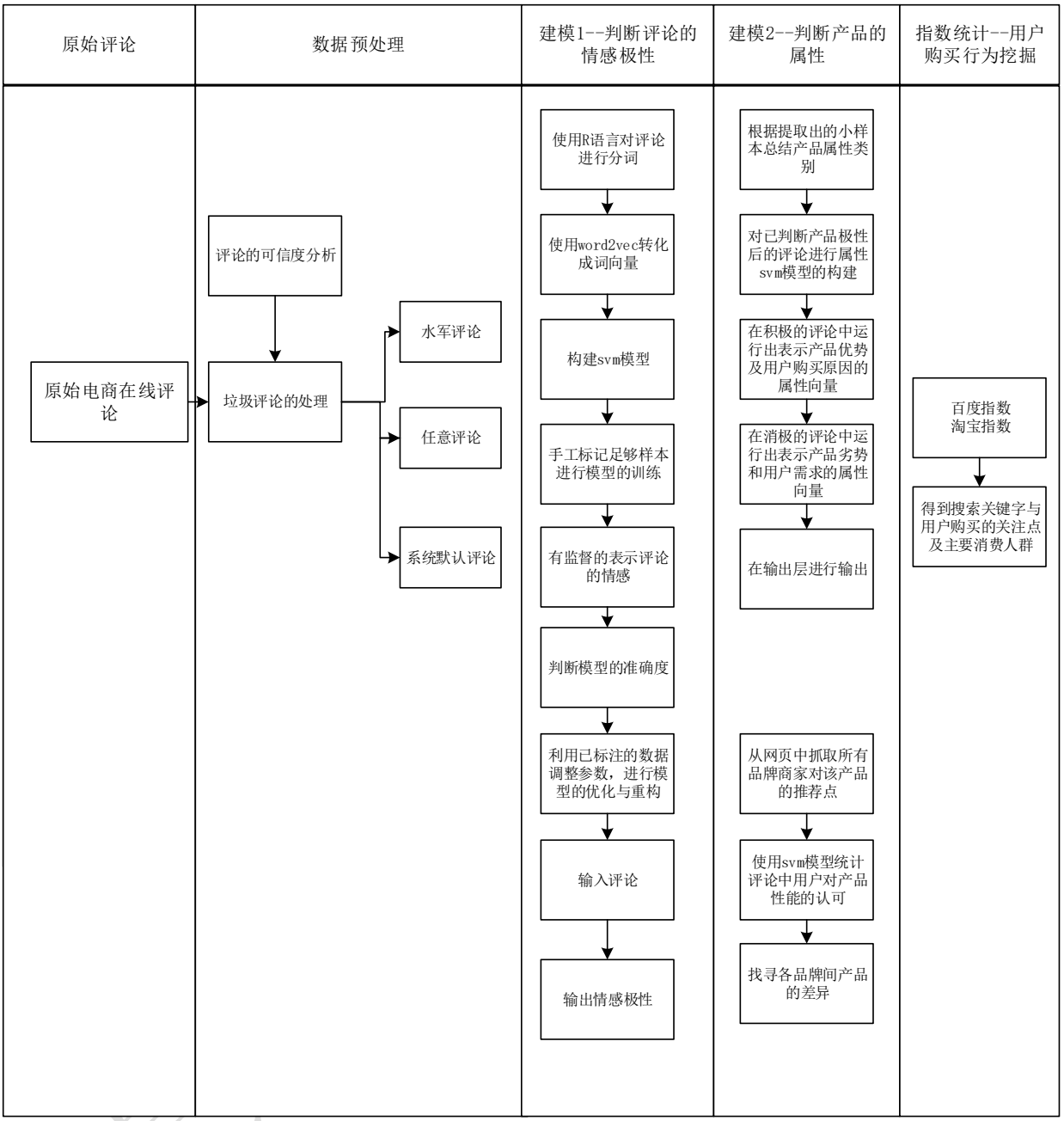

图 1 建模方法及过程的总流程图

2.2 具体步骤

2.2.1 使用火车浏览器对题目所涉及的各品牌产品进行评论的爬取。

2.2.2 对垃圾评论进行处理：

- 垃圾评论的定义：垃圾评论是指那些为了促销某种商品而给出的一些不实际不相符的积极评论，或是为了诋毁某种品牌而给出的一些虚假的负面评论，试图故意误导阅读的人或自动的数据挖掘和情感分析系统的：“不合法”的活动。[1]
- 垃圾评论的分类：对数据进行预处理，根据垃圾评论的识别，将垃圾评论分为以下几种：

  a. 无意义信息，即用户发布的单纯宣泄自己感情的语句，内容空洞，并没有对产品的特征进行分析和评价。

  b. 系统评论，即系统自身默认给出的评论。

c. 评论内容过短，即用户并不是出于对产品进行认真评价的目的进行评价，而是为了网站的积分赠送或者商家的优惠进行的敷衍的评价。

- 在使用编程的过程中设立了几个规则进行作为删除垃圾评论的依据：

a. 由从各大电商网站抓取的评论可得如果用户未作出评论，系统会说默认好评，据此制定规则 1，如果评论中含系统默认好评的，则删除该评论。

b.由于认定评论内容过短无法包含实质信息，因此删除字符串个数小于 6 的评论，据此制定股则 2，如果评论中含字符串个数小于 6 的，则删除该评论。

c.再次对抓取的评论进行分析发现，无意义信息中还包含只有字母或字母个数较多的以及符号或数字过多的评论，因此设定规则 3，如果一条评论中字母的总数/这条评论的长度大于 1/2，或者一条评论中数字的总数/这条评论的长度大于 1/2，则认为该条评论是垃圾评论，删除该评论。

d.由于存在网络延时或者用户重复评论等原因出现的重复评论也是垃圾评论，则据此制定规则，在上下两行数据中若连续两行数据都相等，则删除其中一行数据。

e. 数据预处理后，特征明显的垃圾评论基本被删除，剩下的评论中还包含的垃圾评论为单纯多次宣泄自己感情的语句，也认为此为垃圾评论，使用 R 语言对所有评论进行切割，并统计切割出的词的数量，若评论中正负面评论词的数量/评论中所有评论词的数量大于 1/3，则认为该评论也是垃圾评论，予以删除。

至此，水军、随意发表的评论和默认评论都被删除，垃圾评论基本处理完毕。

2.2.3 使用R语言进行切词。

在R语言中，有一个中文分词效果较好的jiebaR包，将题目给出的excel文件导成csv文件后使用jiebaR进行切词，以热水器为例，将给出的五大电商平台的评论数据全部进行切词处理，最后得到了5个包含所有评论小词块的文本文件。[2] 切词后的文件见附件1.

2.2.4 使用word2vec进行向量化。

将电商平台所有文本文件综合成一份文件进行向量化处理[3]，得到每个单个词块所分配得到的词向量，分配词向量后的文件见附件2.

2.2.5 SVM情感极性模型的构建

在进行情感分析时，考虑一个评论的情感极性，一般来说主要有四个方法：

（1）基于词典的方法

基于词典的方法是将情感词表与人工制定的规则相结合。此类方法最常遇到的问题是常面临无法解决未登录词的问题。

基于词典的方法最简单的做法是构建已知的情感词典，然后遍历需判断的文本，去看文本中包含正向情感词和负向情感词的个数，根据以下公式判断文本的情感极性。

$$Polarity=\begin{cases} Positive(\text{poscnt} > \text{negcnt}) \\ \text{Negative}(\text{poscnt} < \text{negcnt}) \\ \text{Neutral}(\text{poscnt} = \text{negcnt}) \end{cases}$$

词典模型对于本题模型来说最大的劣势在于它需要遍历分词中的每个词及情感词典中每个词才能得到，而对于本题来说数据量过于庞大；另一方面是电商平台上语言规则不统一，种类多样，有些过于通俗而难以被情感词典收录，因此基于词典的统计方法在大量在线电商平台的评论这种方案耗时费力，可行度较低。

（2）无监督的方法

使用 excellent 和 poor 两个种子词与未知词在搜索网页中的互信息来计算未知词的情感极性，并用用于计算整个文本的情感极性。但是在中文中很少有表意与英文一致的句子，即中文中表达相同的情感有很多不同的词，而且像不错与还行这类的词，很难确切得到哪个情感更倾向积极，因此这种确定方法在中文的评论情感分析中没有很适用。[4]

（3）半监督的机器学习方法

半监督模型中较为可行的是基于 RAE 的深度学习模型，但若单独使用 RAE 模型需要在 R 里切割评论，在 Linux 中将词转化为词向量，在 matlab 里使用工具箱，还要在 java 中封装词向量，经过实践发现可操作性很差。

（4）有监督的机器学习方法

有监督的机器学习方法中较为可行的是基于 SVM 模型。

a. SVM 模型简介

SVM 模型是从现行可分情况下的左右分类面发展而来的。在分成两类的平面空间中找寻最优分类线，就是要求分类线不但能将两类正确地分开，使训练错误率为 0，而且还要使分类间隔最大。将其推广到高纬空间，最优分类线就成为最优分类面了。

设 $(x_i, y_i), i=1,...,n, x_i \in R^d$，$y_i \in \{+1,-1\}$ 为两类线性可分的样本集合，对应的线性判别 函数的一般形式为 $f(\mathrm{x})=\omega\square\mathrm{x}+\mathrm{b}$，对应的分类方程如下：

$$\omega\square x+b=0$$

将判别函数进行归一化，使所有样本都满足 $|f(\mathrm{x})|\geq 1$，此时离分类面最近的样本 $f(\mathrm{x})=1$ ，要求分类面对所有样本都能正确分类，既满足

$$y_i[(\omega\square\mathrm{x}_i)+\mathrm{b}]-1\geq 0, \mathrm{i}=1,...,\mathrm{n} \quad \cdots\cdots(1)$$

此时分类间隔等于 $2/\|\omega\|$ ，间隔最大等价于 $\|\omega\|^2$ 最小。最优分类线 H 就是满足上式且使 $\frac{\|\omega\|^2}{2}$ 最小的分类面。

两类数据样本中离分类面最近的样本，且平行于分类面 H 和超平面 H_1、H_2 上的数据样本就是是的上式中等号成立的那些数据样本，这些数据样本就叫做支持向量。

于是将最优分类面问题表示为约束优化的问题，即在式（1）的约束下，求如下函数的最小值

$$\varphi(\omega)=\frac{\|\omega\|^2}{2}$$

之后定义Lagrange函数，并对各个参数求偏导数，且偏导数为0可以将上述最优分类面的求解问题转化为凸二次规划寻优的对偶问题，最终可得最优分类面函数为：

$$f(\mathrm{x})=\mathrm{sgn}\{(\omega\square\mathrm{x})+\mathrm{b}\}=\mathrm{sgn}\{\sum_{i=1}^{n}\partial_i^* y_i(x_i\square x)+b^*\}$$

若 $f(\mathrm{x})=1$，x 就属于该类，否则不属于该类。【5】

b. 评论句子向量化

用word2vec将切词后的词块转化为词向量，借用RAE模型中的递归自编码，将一个句子中的所有词块安装得到父节点，进而将词向量变为句向量，我们利用word2vec获得的是200维向量，举前10维为例子。

具体方法为：

例如，“很”（0.163716 0.038852 0.028363 -0.003588 -0.318706 -0.143267 -0.189193 0.057249 -0.048510 -0.077780…）

“不错”（ 0.098371 0.032743 0.002135 -0.102025 -0.072881 0.013687 -0.041679 0.127724 0.006159 0.112783…）

那么为了表示“很不错”，就假定“很不错”是“很”、“不错”的父节点 p，“很”是第一个子节点 c_1，“不错”是第二个子节点 c_2，那么 p 可由函数 f 从 c_1、c_2 映射得到，如式（2）

所示：

$$p = f(\omega^{(1)}[c_1 : c_2] + b^{(1)}) \quad \cdots\cdots (2)$$

使用贪心算法即每次选取两个节点结合之后父节点得分最高的节点组合，最后可以得到整个句子的 n 维向量表示。[6]

c. 手工标注 400 个句子的情感倾向，训练 SVM 模型。

在已经向量化的句子中判断句子的感情倾向，因为好评居多，因此我们的感情倾向仅分为好评与非好评，不做细分。

d. 将剩下向量化的句子输入 SVM 模型，判断所有热水器评论的感情倾向。

最终结果见输出的 matlab 矩阵。根据矩阵可得好评率为 99.72%，使用程序测试可得使用 SVM 模型获得的准确率为 85%。

2.2.6 产品属性模型的构建

以国美的电商平台为例，对国美中出售的物种品牌的电热水器分别进行词频统计，再对国美平台上的所有热水器汇总进行词频统计。统计结果见图 2.

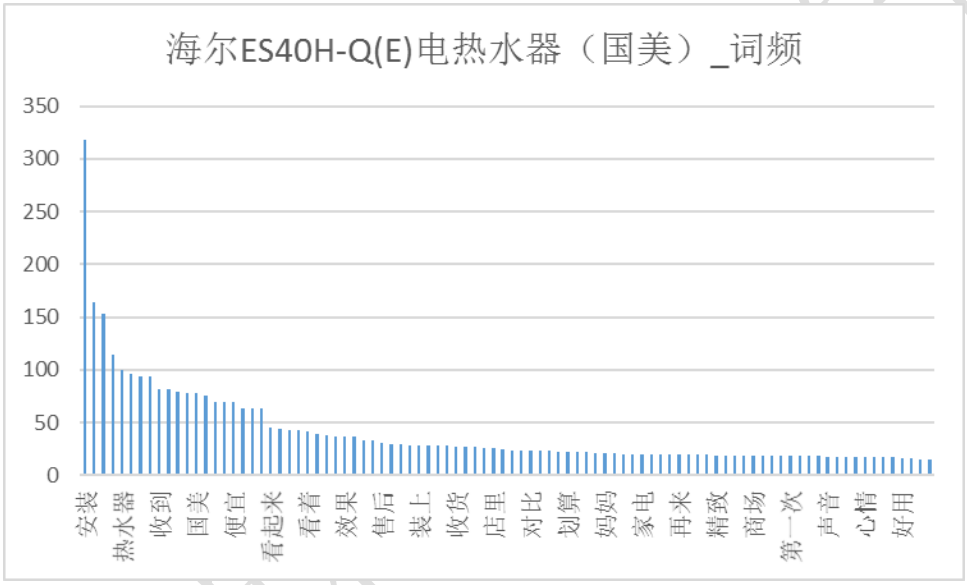

图 2 海尔公司评论的词频统计

从海尔的词频统计来看，安装所占的比例远大于其他词语，可见安装是热水器售后的一项重要服务；其他比较重要的词有便宜划算，说明价格低廉是海尔热水器的一个优势；此外，看起来，看着，效果，这类词词频也比较高，这反映了海尔产品的外观较突出。海尔电热水器可以以价格和外观作为其差异化卖点。

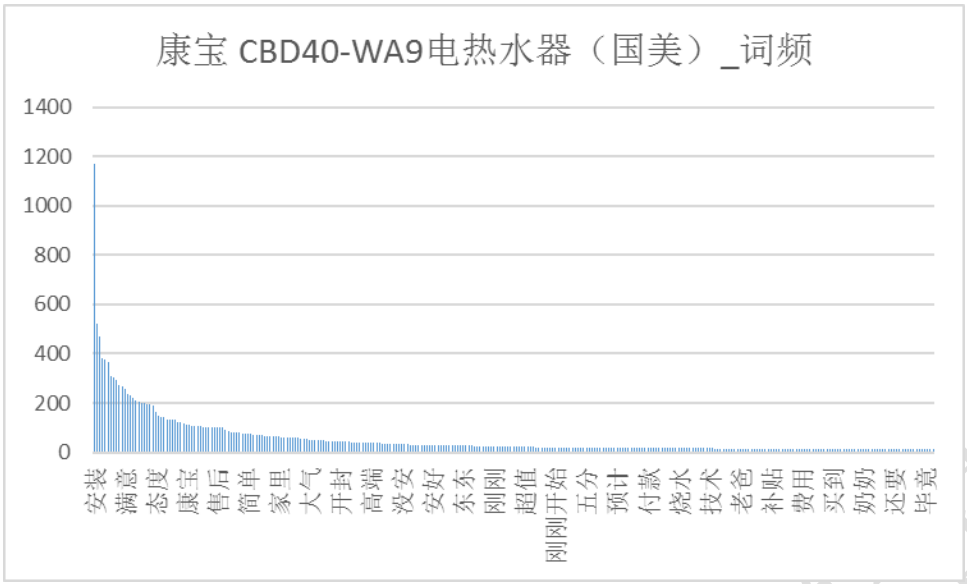

图 3 国美平台上康宝电热水器词频统计

康宝系列提到最多的也是安装问题，接着是售后和服务态度，一个重要的特点是高端大气，外观特点突出。但其他词频并没有很明显比例，可以看到一些用户是卖给自己的亲人用的，刚刚使用不好评价。作为质量和品牌并不突出的电热水器，康宝可以通过改善自己的外观设计吸引更多的客户。

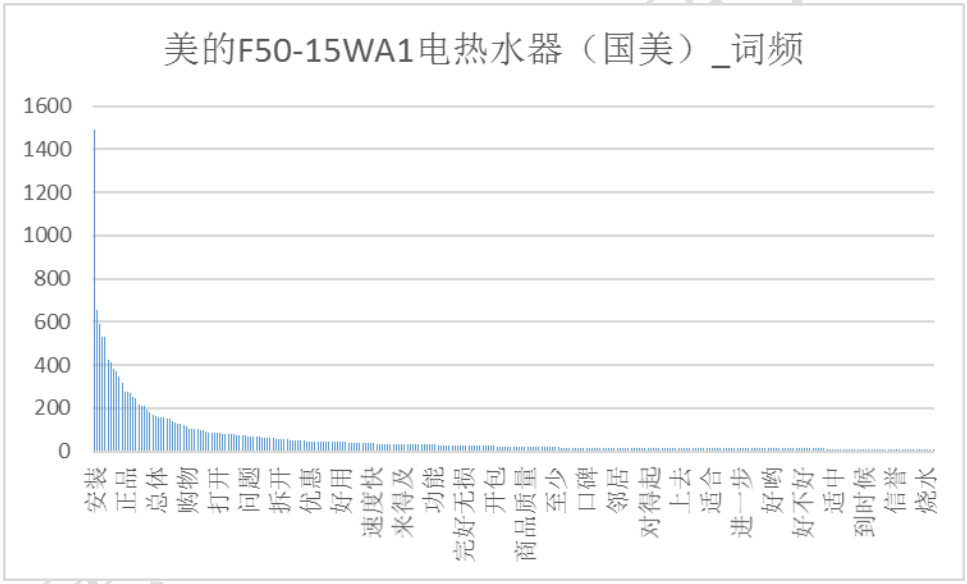

图 4 国美平台上美的电热水器词频统计

美的词频最高的也是安装，用户强调较多的是正品，质量，除去一些不相关的词，可以看出物流，优惠，口碑也是美的热水器的一些卖点。美的作为知名品牌，如果搞一些促销优惠活动，它在大众中的口碑将会增加其销量。

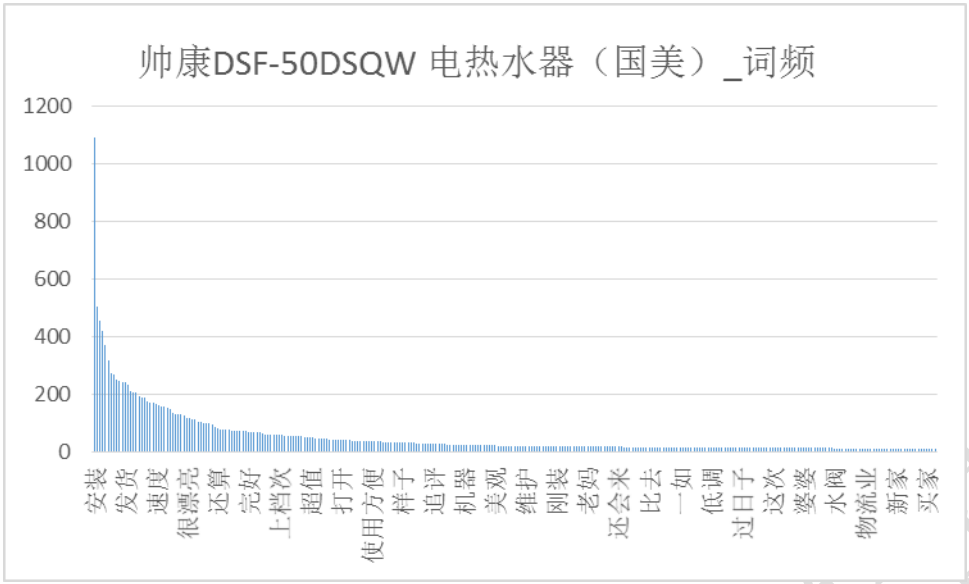

图 5 国美平台上帅康电热水器的词频统计

帅康电热水器客户提到最多的也是安装，其次物流，从发货，速度等词可以看出；从很漂亮，上档次等词可以看出帅康热水器的外观特点也是比较突出。因此，外观也是帅康电热水器的一个重要卖点。

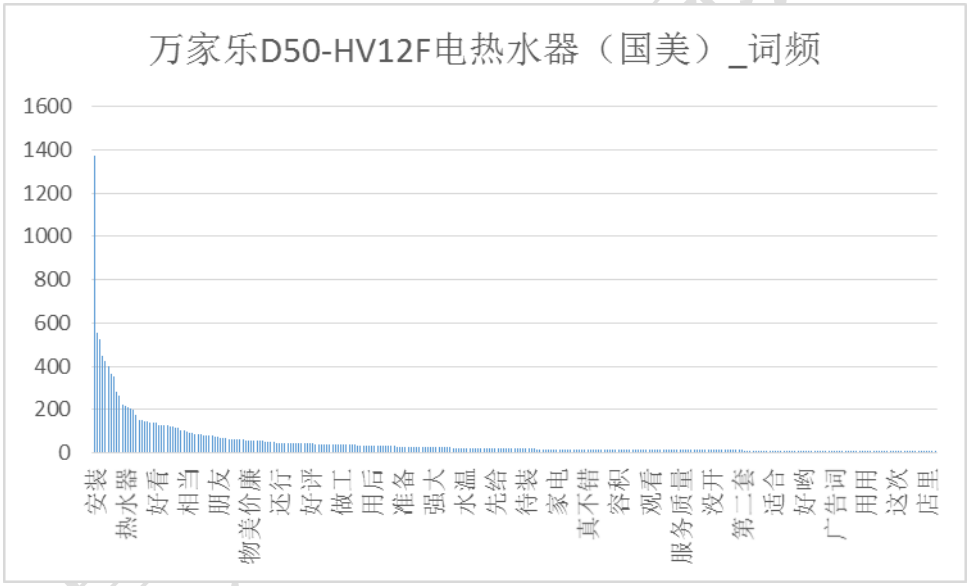

图 6 国美平台上万家乐词频统计

万家乐电热水器安装所占比例较高，有一个突出的特点还是物廉价美，其他关于质量外观的的评论出现相对较少，可能是这种热水器本身的市场占有率并不高，受关注度较少的缘故。作为知名度不高产品，物廉价美是万家乐一个重要卖点，价格上的优势有利于它在市场占一席之位。

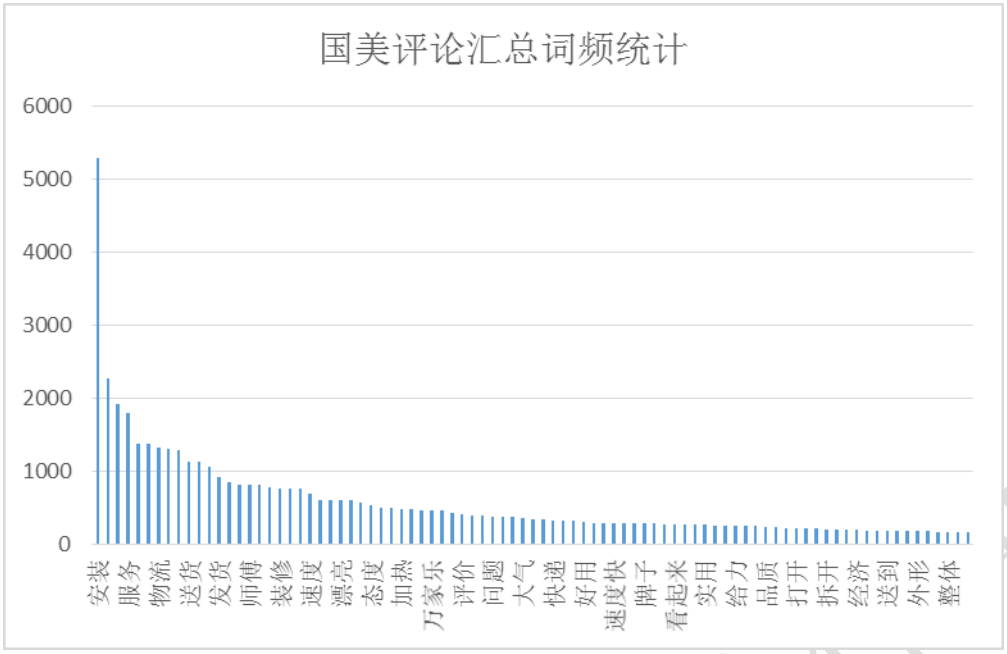

图 7 国美平台上热水器评论汇总词频统计

所以综合各品牌的热水器的特点，可以总结出安装，服务，物流，质量，品牌，价格，外观，性能这八个特点。其中受关注度最高的是安装，注意到师傅这个词的词频也相当高，因此售后服务是客户关注的一个重要部分；可以改善用户体验的有服务、物流、价格、外观这几个点；当一种产品形成品牌效应时，它便能吸引更多的用户。产品的质量和性能并没在评论中有过度的体现，通常是以差评形式出现的，因为大部分热水器在用户评论时间都不会出现太大的问题，需要主意的是，质量性、能很大程度体现在产品的品牌效应上。想打造品牌效应的企业首先要提高自身产品的质量性能。

2.2.7 分析百度指数与淘宝指数。

在百度指数与淘宝指数中分别键入热水器和净水器，根据淘宝指数与百度指数提供的数据与图表，可以进行消费者人群画像，也能够得到用户搜索关键字即购买关注点。

2.3 结果分析

2.3.1 垃圾评论处理的结果分析

通过对垃圾评论的处理，将从各大电商网站中提取到的重复数据及无用数据做了比较彻底地筛选，保证留下的评论数据真实有效，能够从中提取有效信息，为后续进行评论者情感极性的结果分析打下坚实的基础。

2.3.2 SVM 判断情感极性的结果分析

通过手工标识部分评论作为判断情感极性的 SVM 模型的训练语料库，对剩余评论进行句向量的转化后放入 SVM 模型中去，得到约 99%都是好评，因此很难从中直接获得产品的优势和产品的劣势，但是能判断情感极性的正负，在附件中的 matlab 文件中输出了情感极性矩阵，其中情感极性为正则用 1 来表示，情感极性为负的则用-1 来表示。

对 5 个网站的所有评论都进行感情分析，结果如见表 1：

表 1 评论情感极性所占比重

名称	积极情绪	消极情绪	测试评论数
分析结果	99.72%	0.28%	84705

可见差评所占比率并不高。因此分析产品劣势时候可以从消极情感的差评人工提取。

2.3.3 产品属性模型的结果分析

因为消极情感的差评可以人工提取，因此随机抽取部分评论检验模型的准确性，结果见表 2。

表 2 随机抽取检验模型的评论

好评	非好评
海尔的售后真心不错热水器也很漂亮赞一个 速度挺快货已收到海尔值得信赖 用了一段时间才来评价的很好用我很喜欢 装修囤货中还没安装相信海尔的质量 送货快送货人员态度好还会在国美买东西赞 性价比比较高比实体店还便宜 100 大洋呢 国美便宜海尔的品质值得信赖强强联手 安装师傅很负责任产品品质也很好 反正比店里便宜用着吧还挺好的 性价比高信赖海尔	网购比较大的家电还是选了国美没让我失望 便宜是便宜没配件嘛烦 买错了没有半胆速热 还不错就是不知道用久了质量好不好 带遥控器随时加热 卖家到款二天就到东西一样遗憾的是没喷头 超值的热水器带预约带遥控很方便 不停的滴答水烦死啦 发货速度很慢安装时收了我 20 元卸旧热水器的钱 加热时间有点久 就是买完就掉价了特别郁闷

发现除了差评部分有些出入（并不算真正意义上的差评），其他效果都较好。

对国美网站截取的海尔，美的，帅康，康宝，万家乐热水器评论进行感情分析，并提取其中的非好评评论。提取结果见表 3.

表 3 对国美平台上截取各品牌热水器的非好评评论

海尔	美的	帅康	康宝	万家乐
便宜是便宜没配件嘛烦 买错了没有半胆速热 卖家到款二天就到东西一样遗憾的是没喷头 不停的滴答水烦死啦 发货速度很慢安装时收了我 20 元卸旧热水器的钱	还没安装不过这价格实在让我郁闷刚买就降价 还没用温度显示表就掉出来了热水器超级垃圾 快递不好不送到家 买成 779 买后就降价 699 伤心 美中不足的是其实并不怎么省电	到货安装都很好只是感觉加热不快 装好使用了一下水也很烫 东西还可以就是安装费收了我元郁闷 产品不能延迟发货不好 服务网点太少担心售后服务跟不上	产品不错就是没有安装比较麻烦 还好就是过后居然又降价了不爽 安装要求换这换那收费 80 元不爽 装好使用了一下水很烫 还没使用感觉容量不大	到货安装都很好只是感觉加热不快 居然不送货上门害我自己去取还扎爆了轮胎 可惜我们这儿没有售后点要自己安装 装好使用了一下水很烫

发现一个问题，不同品牌之间的非好评部分相同，而且在处理也发现相当多部分的评论是重复的，这可能是题目给出的数据有误。

根据非好评的截取结果，可以看出，非好评数量美的>海尔>康宝>帅康>万家乐。

美的海尔这样的大品牌反而非好评较多，可能的原因有大品牌多人购买，出现的问题也可能增加；也可能是题目提供的资料中大品牌的评论较多，非好评也较多，结果见表 4.

表 4 各品牌产品劣势总结

美的	降价；加热慢；降温快；货损；安装费高；售后差；遥控差；没发票
海尔	没配件；没喷头；滴水；出水量不大；功率抵；安装慢；
康宝	加热慢；烫；降价快；配件贵，售后差；外观丑；
帅康	加热慢；水很烫；物流慢；售后差；配件差；耗电高；
万家乐	加热慢；水很烫；售后差；物流慢

非好评只代表部分客户的意见，并不能反映所有情况，但这些情况应该引起供应商和生产商的重视，客户的负面反馈都可能成为这一产品威胁。

2.3.4 百度指数与淘宝指数对用户购物行为的

1.百度指数显示近 30 天内与热水器相关检索词分别为：

- ✓ 燃气热水器
- ✓ 太阳能热水器
- ✓ 空气能热水器
- ✓ 热水器品牌排行榜
- ✓ 史密斯热水器
- ✓ 热水器什么牌子好
- ✓ 热水器打不着火的原因
- ✓ 即热式热水器
- ✓ 热水器安装

据此检索词排行和百度指数提供的需求分布可得用户在购买热水器时会较为关注热水器的品牌、动力来源、热水器的安全性能以及热水器的能耗以及热水器的售后服务。见**图 8.**

图 8 关键词为热水器的需求分布图

根据百度指数进行的搜索热水器的人群特征统计，可以得出热水器的搜索关注人群年龄在 30 至 39 岁左右，其次为 20 至 29 左右，多为较为年轻的人群，从性别比例来看，男性消费者比例约占 91%，在对热水器的搜索中占了绝对多数。见**图 9**

图 9 购买热水器人群年龄及性别分布图

2. 根据淘宝指数，搜索热水器的买家多为初级消费者，即并不经常在网站上购物，且其消费水平中等的占比最多，达到 53. 8%，由此可得购买热水器的人还是需要一定的消费水平。通过淘宝指数对热水器类目的市场细分的分布进行分析可得用户对电热水器的搜索所占比重最多，达到 63. 82%，其次是燃气热水器和即热式热水器，分别占比为 63. 82%和 5. 51%. 见**图 10**

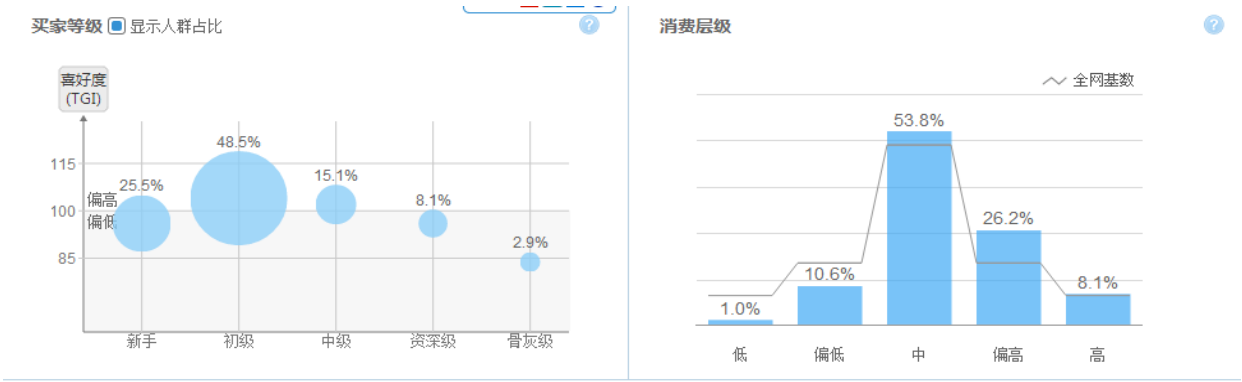

图 10 购买热水器的消费者消费水平分布图

为了更好地分析各品牌与用户搜索量之间的关系，我们从淘宝指数中导出了从 2011/7/11 至今的各品牌用户搜索次数，见散点图，分析后可得排名前三位的品牌分别为海尔、美的和史密斯，且这三种品牌包揽了热水器销售市场的绝大多数份额

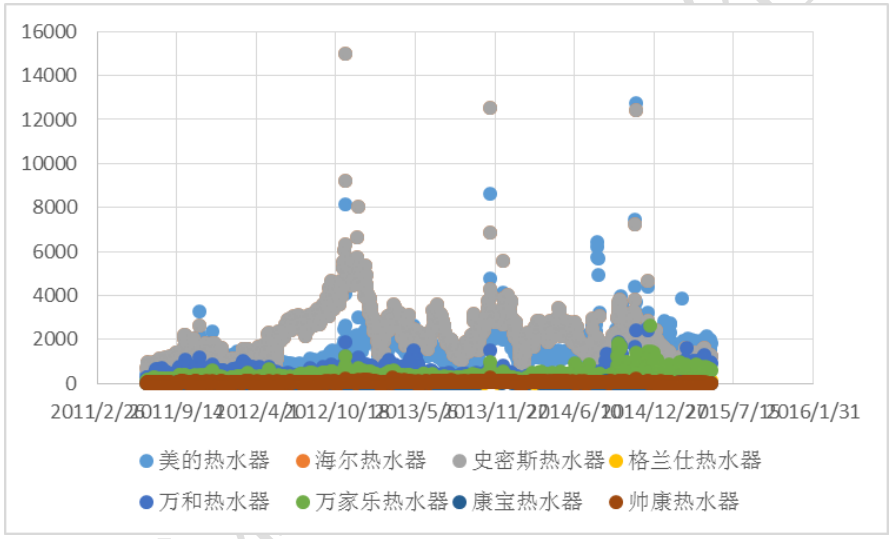

图 11 各大品牌热水器搜索次数近三年来的散点图

据图可得近三年来最受关注的品牌分别为美的、海尔和史密斯。

3. 百度指数净水器需求变化图见图 12：

图 12 关键词为净水器时百度指数需求分布图

从百度指数可以看到，排名牌子等于其关联度较大，可以推断用户对于净水器的品牌较为关注或者

是净水器在品牌树立宣传方面较弱。

4. 淘宝指数分析情况

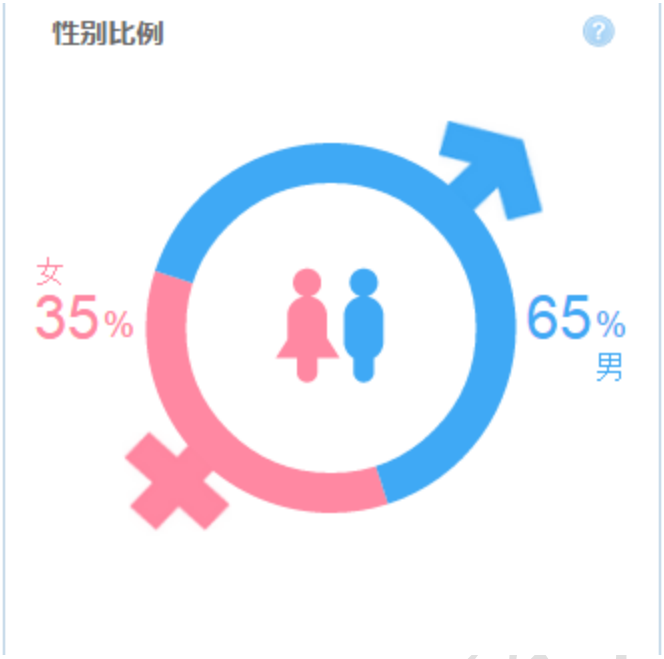

图 13 购买净水器性别比例分布图

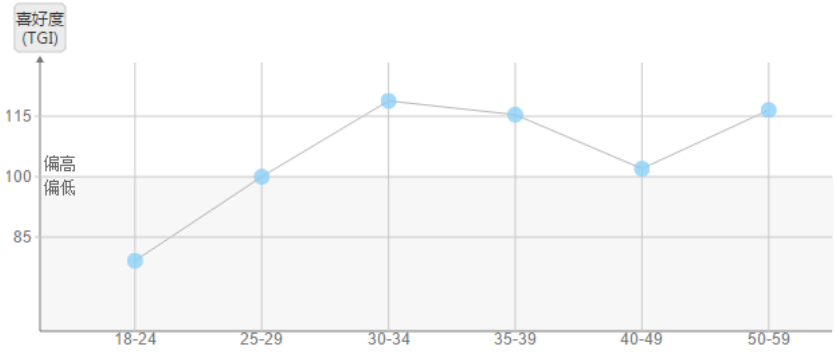

图 14 购买净水器的年龄分布图

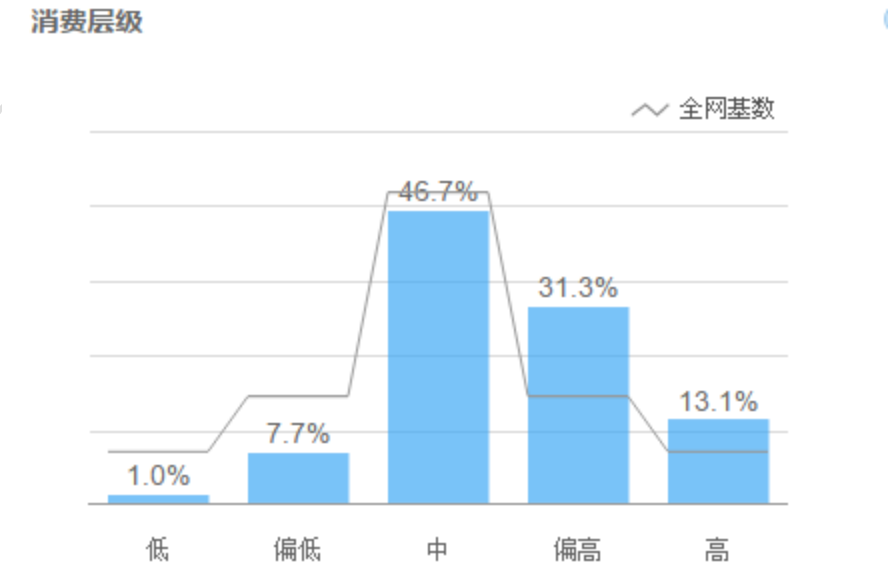

图 15 购买净水器的消费者消费层级分析图

从淘宝指数来看，消费人群男性居多以中老年人士为主，从消费层级可以看到净水器属于中偏高层消费，可以推断消费单位主要以家庭为主，一般生活较为稳定，追求生活品质的人群会更具购买意愿。

3. 结论

如何有效科学利用在线商品评论是文本分析一个重要课题。传统的浅层机器学习的方法是根据感情词典来给分词后的评论进行打分，词典的每一个名词都带有一定的感情分数，通过遍历词典对整个句子的得分进行加总，这种方法需要构造相应的感情词典，但对句子语义多样性并没有很好的解决办法。

本文采用深度学习方法，通过将底层特征进行组合，形成更加抽象的较高层的表示形式，词向量，然后对句子进行训练，然后判断感情极性。考虑到深度学习虽然在训练过程中计算量较大，但是它能够更好的刻画出样本中丰富的内在信息，而且能够避免过拟合的问题。

各大电商平台均是用户的情感极性为正的比例远远大于情感极性为负的比例，这其中可能存在商家为了推销产品篡改评论的行为，但本文就实际数据进行分析就认为事实如此，即用户对各电商平台提供的产品和服务相对是比较满意的。通过对各品牌产品优劣的对比，可以看出各大厂家虽然卖的产品基本相同，但其主打卖点却有着很大不同，其目标群体必然也不同。

通过对百度指数与淘宝指数对用户购物行为的分析，可以看出购买类似电子类产品时一般男性居多，且较为年轻，而购买净水器之类非生活必须品则是由一定消费能力的中年群体为主。

4.参考文献

[1]丁晟春蔡骅.在线评论信息挖掘研究[M].北京：科学出版社，2014:151-154

[2]李明.R 语言与网站分析[M].北京：机械工业出版社，2014:379-387

[3]Word2vec 使用指导
http://blog.csdn.net/jjl2345jjl98999/article/details/11069485

[4]谢丽星.基于 SVM 的中文微博情感分析的研究[D].北京，11-13

[5]丁晟春蔡骅.在线评论信息挖掘研究[M].北京：科学出版社，2014:32-34

[6]梁军，柴玉梅等.基于深度学习的微博情感分析[J].中国信息学报,2014，28(5)