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：一等奖

作品单位：华南师范大学

作品成员：赵晓荣　叶呈成　黄佳锋

指导老师：薛云

基于深度学习的电热水器评论数据挖掘分析

摘要：近年来，随着互联网的广泛应用和电子商务的迅速发展，网络文本及用户评论分析意义日益凸显，因此网络文本挖掘及网络文本情感分析技术应运而生，通过对文本或者用户评论的情感分析，企业能够进行更有效的管理等。本文针对电商平台的电热水器的评论数据，利用基于半监督递归自编码（RAE）的深度学习模型，进行评论的情感分析。为了保证评论数据挖掘分析的质量和全面性，我们重新从京东和苏宁易购平台爬取了评论数据集，对数据进行预处理——评论“去空、去重”、中文分词、停用词过滤等，再利用半监督 RAE 深度学习模型对这些评论进行情感分析。之后，本文主要进行两个方面的数据挖掘分析工作：一方面是根据不同品牌电热水器的评论数据情感分析结果，提炼出各个品牌产品的差异化卖点；另一方面是根据不同电商平台的评论数据情感分析结果，进行不同电商平台的服务质量比较，进而可以使电商平台根据自身优势吸引消费者。

关键词：深度学习，情感分析，RAE，差异化卖点

Data Mining on Comments of Electric water heater Based on Deep Learning

Abstract: Recently, with the wide application of Internet and the rapid development of electronic commerce, network text and user review analysis is of great significance, text mining and sentiment analysis of network text arise at the historic moment, and the emotional analysis of the text or user comments is more effective in enterprise management and so on. Electric business platform, this paper apply a deep learning method based on semi-supervised recursive encoding (RAE) on analysis of the emotion of comments which users delivered about electric water heater. In order to ensure the quality of the data mining analysis, we crawled the relevant comments data sets from Jingdong and Suning platform. Then we preprocessed comments data on wiping "empty and heavy" out, Chinese word segmentation, filtering stop words, word frequency statistics, etc. Next we analyze sentiment on these comments using a method based on semi-supervised RAE. Later, this paper analyzed mainly comments in two aspects of data mining work: on the one hand, according to sentiment analysis result of the comments of different brand electric water heater, extracting differentiation of various brand products selling point; On the other hand, according to the comments of different electric business platform data sentiment analysis results, and compare different electric business platform of service quality, and electric business platform can take measures to attract consumers according to their own advantages .

Key words: deep learning; sentiment analysis; RAE; differentiation of selling point

目 录

1. 挖掘目标

本次建模针对电商平台上关于电热水器的评论数据，采用基于半监督 RAE 深度学习模型的数据挖掘方法，达到以下两个目标：

1) 利用半监督 RAE 模型对同一品牌电热水器的评论进行情感分析，根据分析结果得到用户针对各属性的满意度，从而提炼出该产品的优势和劣势。分析不同品牌电热水器的评论数据，提炼出其差异化卖点。

2) 对不同电商平台对应相同电热水器的评论数据进行情感分析，根据分析结果得出各个电商平台服务的优势与劣势。

2. 分析方法与过程

2.1. 总体流程

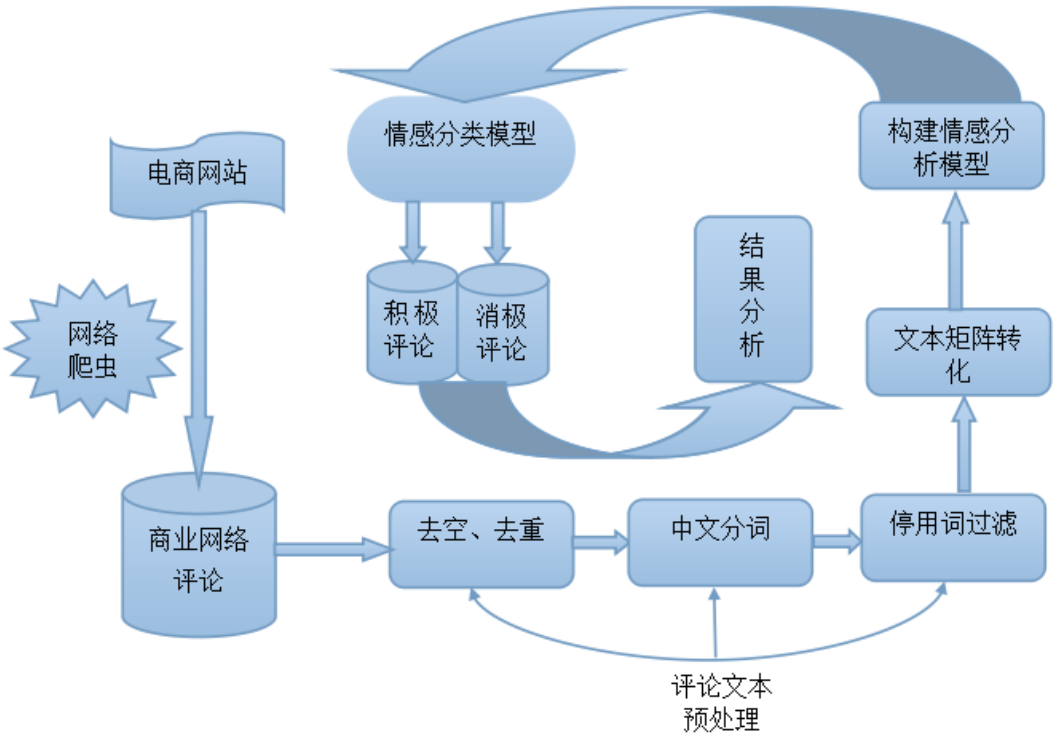


图 1 总体流程图

本用例主要包括以下几个步骤：

步骤一：爬取网络评论数据，评论数据的获取是本次数据挖掘分析的第一步。本文中利用火车头数据采集器，对评论文本进行抽取，最后将评论文本批量存进 txt 文件中，得到实验数据。

步骤二：数据预处理，直接从网上爬取的评论数据中往往不能直接分析需要进行数据预处理。第一步要“去空、去重”；第二步对评论数据进行中文分词，将一句评论分成多个词语进一步分析；第三步进行停用词过滤，去除掉评论中与情感判定不相关的词。

步骤三：文本矩阵转化，使用基于半监督 RAE 深度学习模型进行情感分析，需要将文本词语全部转换为词向量，本论文中构建了一个词表和词向量表，词表中为全部文本词语和词语的编号，词向量表中为全部词语的词向量。

步骤四：情感分析，构建基于半监督 RAE 的深度学习模型，利用选出的积极、消极评论各占一半左右的数据集训练情感分析模型，并进行测试，得到符合要求的模型。利用构建的模型分析得出评论数据的情感倾向。

步骤五：属性提取并统计，将所有提及到电热水器的某些属性的评论数据从实验数据集中筛选出来，统计各个属性相关评论数据的积极评论和消极评论占该产品的积极评论和消极评论的百分比。

步骤六：结果分析，根据分析结果提取产品的差异化卖点或者每个电商平台的竞争优势和劣势，进而制定合适的营销策略。

2.2. 具体步骤

步骤一：爬取网络评论数据

随着电子商务的迅速发展，网购的消费者越来越多，他们不再只是被动的获取网络知识，而是可以通过网络发表产品评论来分享自己的用户体验，而评论中所包含的丰富信息, 对企业管理具有重要的价值。通过数据挖掘等技术手段实现对客户评论的智能分析，商家可以获得客户对产品的意见和态度，获取网络评论数据中的有价值的信息，做出相应的营销策略和产品改进方案等。而网络数据挖掘分析的第一步就是爬取网络评论数据。

本次论文中采用火车头数据采集器爬取网上评论数据，将批量的 URL 存放

进采集队列中，设置采集内容的规则，从评论网页上爬取实验需要的评论文本数据，详细步骤如下：

1）采集网址规则

我们首先采集美的 F50-21W6 的评论数据，打开它的评论页面我们要采集的评论共有 6065 条，分 203 页显示，如图 2 所示：

图 2 美的 F50-21W6 评论页面

为采集该商品的所有评论数据，这里采用批量网址采集，将 203 个网址导入进行数据采集，如图 3 所示：

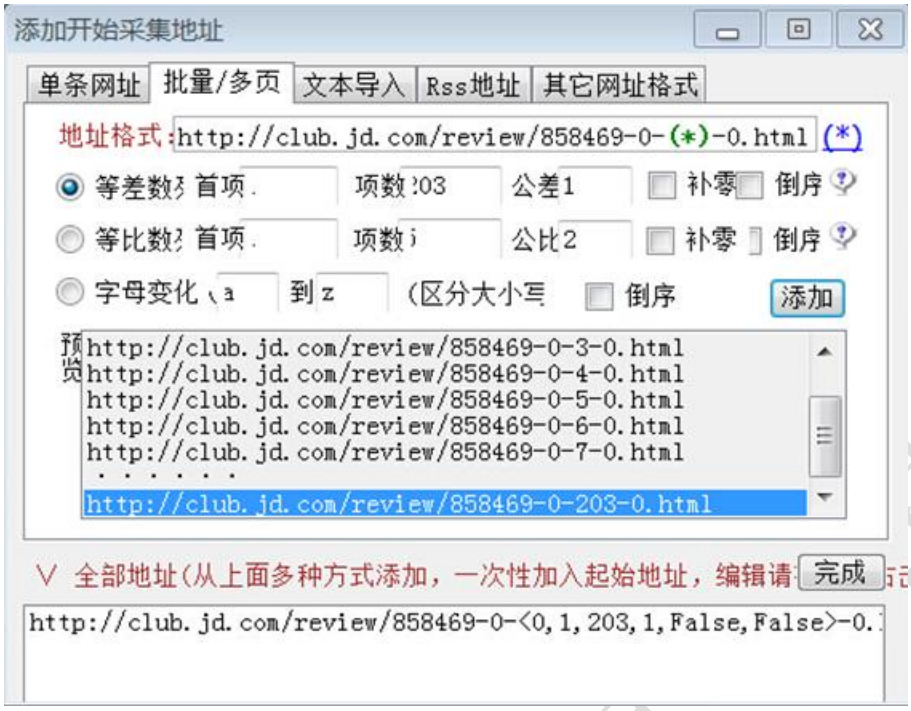


图 3 批量网址采集规则设置

2）设置采集内容规则

为了抽取出网页中有用的网络商业评论信息，还需要对采集内容规则进行设置。首先在京东网上打开美的 F50-21W6 的评论页面，可以看到在京东网上评论的标签为“心得”。接下来打开该页面的源代码，搜索到“心得”部分，可以发现它的结构如下：

```
<dl>
<dt>心　　得：</dt>
<dd>不错！性价比非常高！</dd>
</dl>
```

其中的“不错！性价比非常高！”就是我们想要的网络商业评论文本。最后，根据评论在 HTML 文档中的结构分布，设置采集内容规则，如图 4 所示

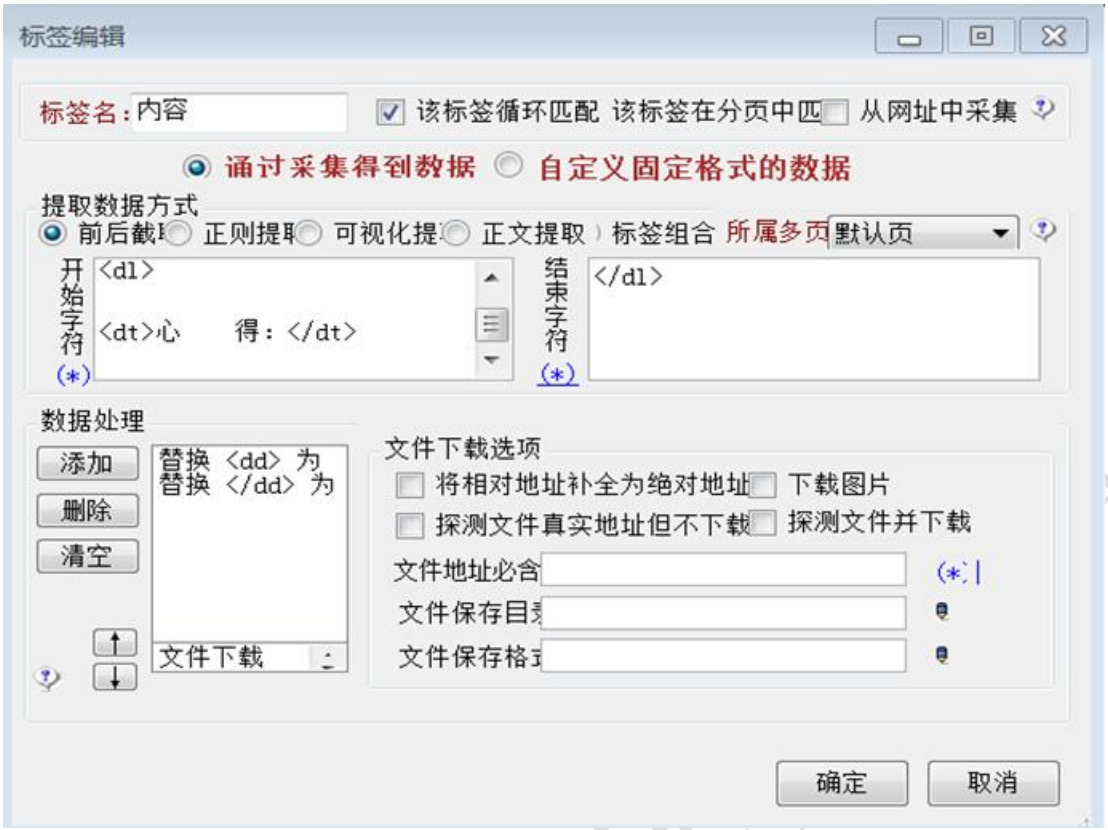


图 4 采集内容规则设置

3）结果发布

为了后续研究工作的方便，本文选择将采集到的网络商业评论存储在同一个 txt 文件中，文件编码为”UTF-8”，最终得到一个存储全部评论文本的 txt 文件。美的 F50-21W6 的评论示例如下：

美的电热水器质量不错，价格比店里要便宜。

物流给力机子不错很好

很好看也很实用，配送很快，安装师傅人也很好的。

头天下单，第二天就到货安装好了，非常满意

本文实验中：从京东上选择了三个品牌的电热水器的评论数据进行抓取——美的 F50-21W6、海尔 EC5002-D、格兰仕 G50E302T，用于提炼不同品牌产品的差异化卖点；从苏宁易购上爬取了美的 F50-21W6 电热水器的评论数据，用于比较和京东电商平台的服务特点。本次实验数据见附件。

步骤二：数据预处理

与数据库中的结构化数据相比，从网页上爬取的数据属于半结构化或者非结构化数据，即具有有限的结构，或者根本就没有结构，即使具有一些结构,也是着重于格式，而非文档内容，不同类型文档的结构也不一致。此外，网页数据缺乏机器可理解的语义，而数据挖掘的对象局限于数据库中的结构化数据，并利用关系表格等存储结构来发现有价值的信息,因此有些数据挖掘技术并不适用于网络文本挖掘，即使可用也需要建立在对网络文本数据进行预处理的基础之上。如果要对网络评论数据进行情感分析，就必须先将文本数据进行预处理，转化为结构化的数据。该步骤中，从以下几个方面对步骤一中从网页上爬取的评论数据进行预处理。

1）“去重”、“去空”

对于存储了全部网络商业评论的 txt 文件，每行代表了一个评论文本但是难免会出现两个完全一样的文本和一些空行。所以本文首先进行了“去重”、“去空”的预处理工作。

在导入评论文本时，同时进行了是否为空的判断，只导入不为空的文本，从而过滤掉了空白文本，“去空”的程序段如图 5 所示：

```
StreamReader sr = new StreamReader("C:/Users/IBM/Desktop/电热水器数据/
京东/F50-21W6.txt", Encoding.UTF8);
String line;
while ((line = sr.ReadLine()) != null)
{
    if (line.ToString() != "")   //去掉空文本
    {
        CommentsList.Add(line.ToString());
    }
}
```

图 5“去空”程序段

将非空的评论文本导进 List 后，再进行去除重复处理，过滤掉重复的评论文本，“去重”的程序段如图 6 所示：

```
CommentsList2.Add(CommentsList[0]);
for (int i = 1; i < CommentsList.Count; i++)
{
    IsRepeated = false;
    for (int j = 0; j < i; j++)
    {
        if (CommentsList[i].Equals(CommentsList[j]))
        {
            IsRepeated = true;
            break;
        }
    }
    if (!IsRepeated)
    {
        CommentsList2.Add(CommentsList[i]);
    }
}
```

图 6“去重”程序段

2) 中文分词

中文分词(Chinese Word Segmentation)，也可称为中文切词，指的是通过某种特定的规则，将中文文本切分成一个一个单独的词。本文使用 NLPIR 汉语分词系统（又名 ICTCLAS 2015）进行分词，它是中科院张华平博士主持开发的中文汉语分词工具，主要功能包括中文分词；词性标注；命名实体识别；用户词典功能；支持 GBK 编码、UTF8 编码、BIG5 编码。新增微博分词、新词发现与关键词提取功能。本文用到了在 NLPIR 官网上下载到的 NLPIR.dll 程序包，在 Microsoft Visual Studio 2012 编程环境中用 C#高级语言程序对 NLPIR.dll C++程序包进行调用，实现对网络商业评论文本进行批量分词处理和词性标注。主要程序段如图 7 所示：

```
if (!NLPIR_Init("F:/ICTCLAS2015", 0, ""))
{
        System.Console.WriteLine("Init ICTCLAS failed!");
        return;
 }
 else
        System.Console.WriteLine("Init ICTCLAS success!");
 Console.WriteLine();
 System.Console.WriteLine("分词处理中...");
  for (int i = 0; i < content.Count; i++)
  {
        /***************分词***************/
        IntPtr intPtr = NLPIR_ParagraphProcess(content[i]);
        String str = Marshal.PtrToStringAnsi(intPtr);
        content_seg.Add(str)
  }
```

图 7 批量中文分词程序段

分词结果示例：

分词前： 物流快！服务好！物品嘉！

分词后：物流/n 快/a ！/wt 服务/v 好/a ！/wt 物品/n 嘉/b ！/wt

从上述结果可以看出，本文已经将网络商业评论文本切分成一个个的词语，“/”后面是对应词语的词性标注（如：“v”代表动词，可对照中科院《计算所汉语词性标记集》）。

3) 停用词过滤

评论文本在经过去重、去空、中文分词后，并非所有的剩下的词语都可以作为特征词，里面还有一些包含的信息量很低甚至没有信息量的词语，需要将它们过滤掉，否则将会影响下文的分析的正确率。在信息检索中，为节省存储空间和

提高搜索效率，在处理自然语言之前会自动过滤掉某些字或词，这些字或词即被称为 Stop Words（停用词）。

本文采用了“词性+停用词表”的过滤方法。在上文已经提到了中文分词后的词语还带有词性的标注，所以本文根据中科院《计算所汉语词性标记集》将上述停用词词性都写进 StopwordPropsList 里面，如图 8 所示，然后对每个分词后的文本进行遍历扫描，把对应词性的词语全部过滤掉。

```
//介词
StopwordPropsList.Add("p");
StopwordPropsList.Add("pba"); //把
StopwordPropsList.Add("pbei"); //被

//连词
 StopwordPropsList.Add("c");
 StopwordPropsList.Add("cc");   //并列连词
```

图 8 停用词词性列表（部分）

为了把评论文本中包含的停用词过滤干净，本文还利用了《哈工大停用词表》进行辅助过滤，在词性过滤后再把文本中存在于停用词表的词语过滤掉，进一步过滤掉评论文本中的停用词。

停用词过滤结果示例：

分词后：第一/m 次/qv 在/p 苏宁/nz 易/ad 购/vg 购买/v ，/wd 购买/v 和/cc 售/v 后/f 都/d 很/d 满意/v ，/wd 不仅/c 优惠/vn 事/n ，/wd 下次/t 继续/v 合/v 又/c 省/n 作/v

停用词过滤后：第一 苏宁 易 购 购买 购买 售 后 都 很 满意 优惠 事 下次 继续 合 省 作

经过上述步骤的数据预处理后，实验数据的数量如下表 1 所示：

表 1 预处理后的评论数据数量

京东美的 F50-21W6	京东海尔 EC5002-D	京东格兰仕 G50E302T	苏宁美的 F50-21W6
1381	1293	1636	2775

步骤三：文本矩阵转化

目前，在文本情感分析中，主要的研究方法还是基于机器学习的方法。如果想利用机器学习的方法进行情感分析，第一步就是要找一种方法将文本数据特征符号数学化，将文本数据转化为计算机可以识别的数字信息。最初的学者利用传统的 One-hot Representation 的方式实现文本矩阵转化，建立一个词库向量维度等于词表大小，某句文本评论中出现某个词语，该词语对应的维度的值为 1，不出现则为 0，用这种方法建立的文本矩阵是一个维数较大且稀疏的向量矩阵，使后面情感分析的计算量大大增加，且准确率不高。本文中是将词语用一个 n 维实数向量去表示，其基本的思想是通过训练，将语料中的词语映射到 n 维实数向量，这种词语的表示方式优于 One-hot Representation 方法，n 维向量不但包含了词语间的潜藏语义关系，同时也避免了维数灾难。Ronan Collobert 和 Jason Weston 于 2008 年推出 S E N N A 系统，使用词向量方法去完成自然语言处理中的各种任务，例如，词性标注、命名实体识别、短语识别、语义角色标注等。本文中也利用词向量的方法将文本数据转化为结构化的向量矩阵，进一步进行情感分析。

1）向量化概述

文本矩阵转化的第一步就是词向量化，顾名思义，词向量化即用空间向量模型表示各个词语，进而提高计算机对自然语言的处理能力。词向量具有良好的语义特性，是表示词语特征的常用方式。情感分析中把对文本内容的处理简化成对一定长度的向量的处理时，通常使用较低维度的空间向量来表示词语的特征，避免数据维数灾难。词向量中每一维的值代表一个具有一定的语义和语法上解释的特征。

词向量化后便可以将评论的文本数据转化向量矩阵了。通常情况下，我们将词语 w 映射到 n 维空间向量，即 $w \in R^n$，一个文本或者句子中含有 m 个词语，把这 m 个 n 维空间向量堆放在一起，就得到整个文本或句子的空间向量模型——一个词向量矩阵 $L \in R^{m \times n}$。例如给定句子 c 含有 m 个词语，$1 \le i \le n$，w_i 为句子 c 的空间向量矩阵 L 中的第 k_i 列，即可 $w_i = Le_{k_i} \in R^n, e_{k_i} \in R^m$，且除了第 k_i 个分量为 1，其余分量

均为 0.

将一个文本或者一句评论映射成一个词向量矩阵后，即将中文文本数据转化成计算机可以识别的信息格式，继而利用基于递归自编码的深度学习方法进行情感分析。

2）文本矩阵转化过程

通过编写程序产生随机的向量词表，每个词对应一个唯一的词标识号和词向量，如图 9 和图 10 所示，例如“认识”的词标号为 3，在词向量表中，列号为 3 对应的列向量便是标识“学会”的词向量。词向量表生成后，通过扫描，将每句评论转化成一个词向量矩阵，将中文文本数据转化成数字数据——计算机可以识别的数据信息，进而进行文本情感分析。此步骤的详细实现程序见附件。

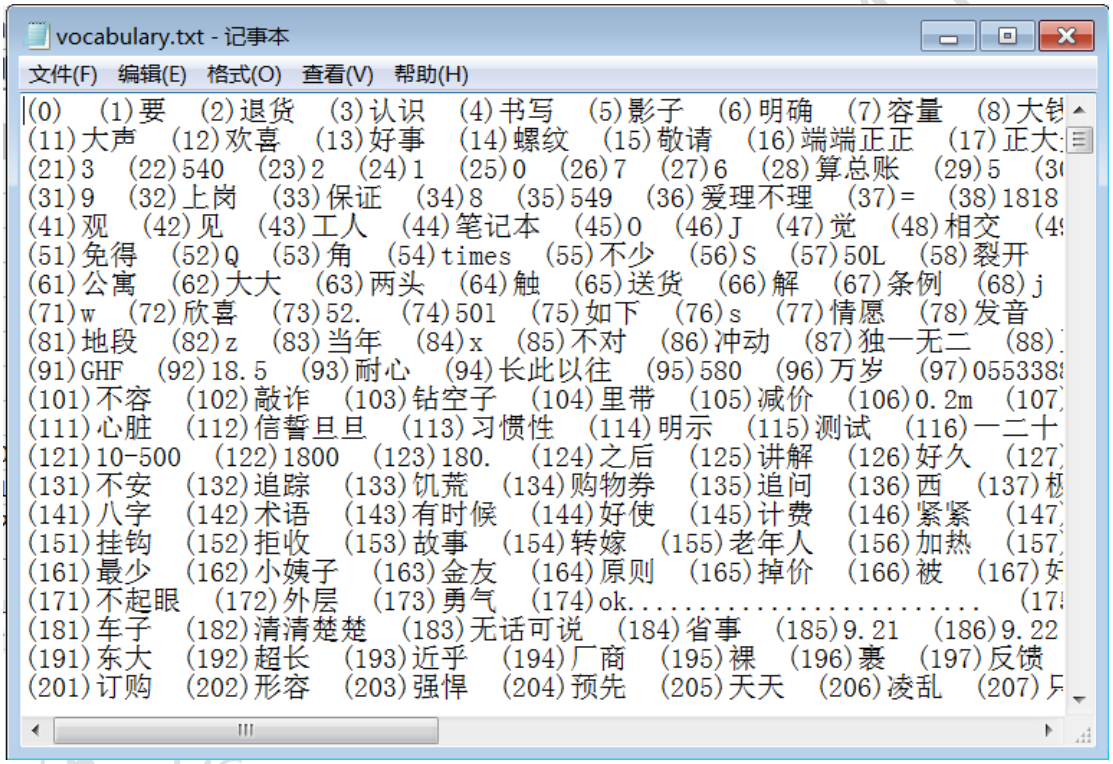
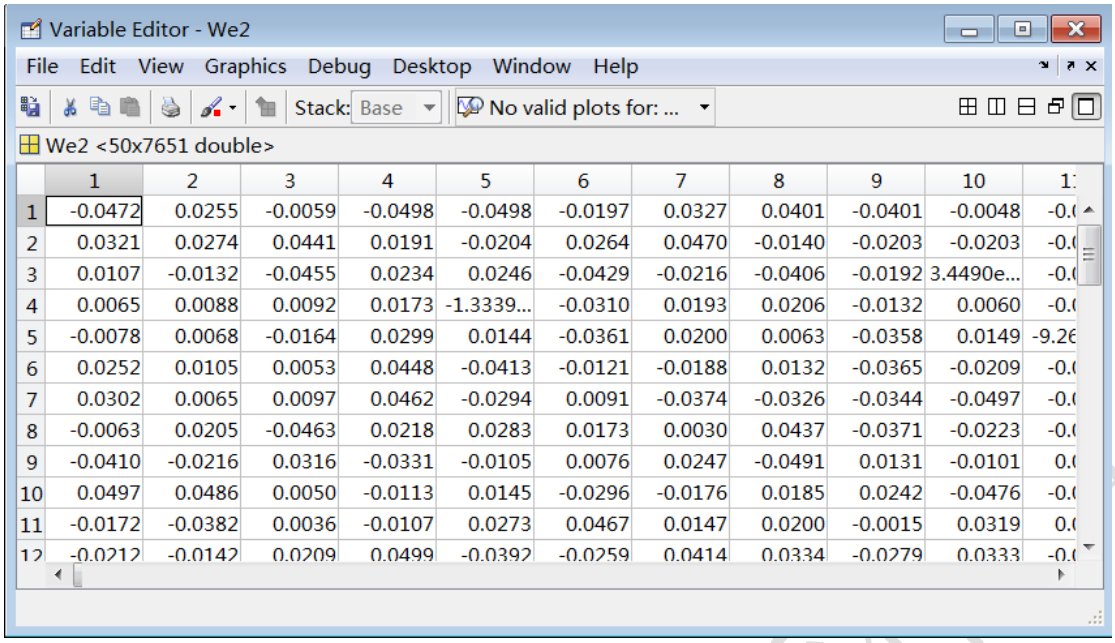


图 9 词表



	1	2	3	4	5	6	7	8	9	10	11
1	-0.0472	0.0255	-0.0059	-0.0498	-0.0498	-0.0197	0.0327	0.0401	-0.0401	-0.0048	-0.0
2	0.0321	0.0274	0.0441	0.0191	-0.0204	0.0264	0.0470	-0.0140	-0.0203	-0.0203	-0.0
3	0.0107	-0.0132	-0.0455	0.0234	0.0246	-0.0429	-0.0216	-0.0406	-0.0192	3.4490e...	-0.0
4	0.0065	0.0088	0.0092	0.0173	-1.3339...	-0.0310	0.0193	0.0206	-0.0132	0.0060	-0.0
5	-0.0078	0.0068	-0.0164	0.0299	0.0144	-0.0361	0.0200	0.0063	-0.0358	0.0149	-9.26
6	0.0252	0.0105	0.0053	0.0448	-0.0413	-0.0121	-0.0188	0.0132	-0.0365	-0.0209	-0.0
7	0.0302	0.0065	0.0097	0.0462	-0.0294	0.0091	-0.0374	-0.0326	-0.0344	-0.0497	-0.0
8	-0.0063	0.0205	-0.0463	0.0218	0.0283	0.0173	0.0030	0.0437	-0.0371	-0.0223	-0.0
9	-0.0410	-0.0216	0.0316	-0.0331	-0.0105	0.0076	0.0247	-0.0491	0.0131	-0.0101	0.0
10	0.0497	0.0486	0.0050	-0.0113	0.0145	-0.0296	-0.0176	0.0185	0.0242	-0.0476	-0.0
11	-0.0172	-0.0382	0.0036	-0.0107	0.0273	0.0467	0.0147	0.0200	-0.0015	0.0319	0.0
12	-0.0212	-0.0142	0.0209	0.0499	-0.0392	-0.0259	0.0414	0.0334	-0.0279	0.0333	-0.0

图 10 词向量表

步骤四：情感分析

情感分析自从 2002 年由 BoPang 提出之后，获得了很大程度的关注，特别是在在线评论的情感倾向性分析上获得了很大的发展。文本情感分类在情感分析研究中占有举足轻重的地位，在信息爆炸的 21 世纪，海量数据的情感分类研究吸引了很多的研究者，如何深入学习文本的语义信息，准确表达语义特征，提高情感分类的准确性是研究的目标。

目前，情感分析的主要研究方法还是一些基于机器学习的传统算法，例如，SVM、信息熵、CRF 等，机器学习的第一次浪潮是浅层学习，深度学习则是机器学习的第二次发展浪潮。以往的情感分析主要是采用浅层学习，但是无法学习文本语义信息，随着技术的发展和科技的进步，人们的要求也随之越来越高。在大数据的分析和处理上浅层学习存在的弊端导致情感分析遇到了瓶颈，因此人们将焦点转移到了可以改善这一弊端的深度学习的研究。2003 年 Bengio 等人提出用神经网络构建二元语言模型的方法；2006 年，机器学习领域的泰斗，加拿大多伦多大学教授 Geoffrey Hinton 和他的学生 Ruslan Salakhutdinov 在《科学》上发表文章，从此开启了在学术界和工业界对深度学习的研究浪潮，他们提出来两个观点：其一，多隐层的人工神经网络具备着优异的学习特征的能力，它学习到的特征对样本数据有着更加本质的刻画，使其更加有利于图像可视化或者文本等的分类任务；其二，深度神经网络在训练的时候存在一定的难度，这些可通过

“逐层初始化”（layer-wise pre-training）的方法来有效的克服掉，在文章中是采用无监督学习来完成逐层初始化的工作的。2006 年，Hinton 等人基于深信度网络(DBN，Deep Belief Nets）提出了非监督学习的贪心逐层训练算法，给解决深层结构中相关的优化难题带来了希望，之后提出了多层自动编码器的深层结构。后来，Lecun 等人采用的是卷积神经网络（CNNs，Convolutional Neural Networks），这是第一个真正具有多层结构的学习算法，它使用空间的相对关系来减少参数数目进而提高 BP 训练性能。2011 年，Socher 提出基于递归自编码器（Recursive AutoEncoder, RAE）的树回归模型用来分析句子的情感倾向性，本文引用 Socher 提出的半监督 RAE 的深度学习模型进行情感分析。

1) 半监督 RAE 的情感分析模型概述

a. 传统的递归自编码（简称 RAE）

传统的递归自编码（简称 RAE）是自编码方法的一个变种，它属于深度学习一种方法，近年来被 Socher 等人应用于情感分析领域，这种深度学习的方法是多隐层的神经网络结构，可以逐层分析，优化每一层学习得到的特征向量表示，因此它抽取的文本特征向量可以更准确的表达语义信息，提高分类结果。

自编码的作用是学习输入数据隐含的特定结构，传统的自编码会对输入给定一个树结构，图 11 表示的就是一个给定的递归自编码的树状结构，此时假设我们给出一个句子的词向量的列表 $x=(x_1,\cdots,\cdots,x_m)$，**错误！未找到引用源。**上一层节点以及二叉树结构的输入用一个包含一个父节点和两个子节点的三元组表示：$(p\to c_1c_2)$。每个子节点可以是一个输入字向量 x_i 或者是树中的非终端节点。以图 11 为例，我们有以下三元组：$((y_1\to x_3x_4),(y_2\to y_1x_2),(y_3\to y_2x_1))$，其中隐层表示 y_i 必须与词向量 x_i 的维度相同。

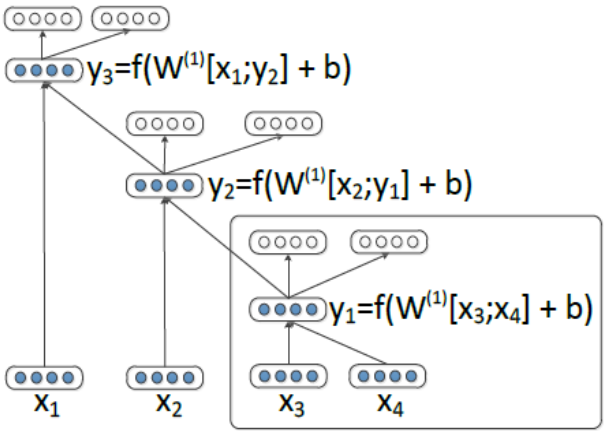


图 11 递归自编码的树结构

从这种树状图中，我们可以计算父节点的表示。这第一个父节点向量 y_i 通过子节点 $(c_1,c_2)=(x_3,x_4)$:

$$p = f(w^{(1)}[c_1;c_2]+b^{(1)}) \quad (1)$$

其中， $w^{(1)} \in R^{n\times 2n}$ 是参数矩阵， $b^{(1)}$ 是偏差，n 为空间向量的维度。我们乘以两个并置子节点参数矩阵 $w^{(1)} \in R^{n\times 2n}$ **错误！未找到引用源。**，加入偏差项之后，我们把每个结果带入函数中如双曲正弦中去评估所得到的向量，此外，通过增加重构层（图中空心部分）重构该父节点的子节点的方式判断得到的父亲节点是否能够很好的表示子节点信息，评估的方法之一就是如何更好的用 n 维向量表示为了重构在重构层的子节点。

$$\left[c_1';c_2' = w^{(2)}p + b^{(2)}\right] \quad (2)$$

训练过程中，目标是最小化重构子节点与原来的子节点之间的误差，即重构误差。图中矩形框中的部分是 RAE 方法中的一次迭代计算，在每次迭代中，采用欧氏距离衡量衡量重构误差，如公式所示

$$E_{\text{rec}}([c_1,c_2]) = \frac{1}{2}\left\|[c_1;c_2]-[c_1';c_2']\right\|^2 \quad (3)$$

至此，一个三元组的向量表示确定，而树形结构中的其他三元组的计算也采

用相同的计算方法，实质上，就是重复上述动作，直至重构误差达到设定的阈值。

b. 基于半监督 RAE 的深度学习模型

传统的 RAE 递推自编码是完全无监督和一般情况下多字词组的语义捕捉，他的一个缺点就是词与词之间没有建立联系。我们扩大传统无监督 RAE 的应用范围到半监督 RAE，引入半监督 RAE 的机制，预测句子或者短语的情感分布。它的核心思想在于计算文章中的交叉熵误差（cross-entropy error）和重构误差（reconstruction error）。

在半监督 RAE 中，在每一个父节点上增加一个简单的 softmax 层，辅助预测类分布：

$$d(p;\theta) = soft\max(w^{label} p) \quad (4)$$

假设有 K 个情感标签，$d \in R^K$ 是 K 维向量分布而且 $\sum_{k=1} d_k = 1$（如果只有两类情感分布：积极和消极，此时便是 2 位向量分布[0,1]或者[1，0]）。图 12，表示的就是一个半监督 RAE 过程，让 t_K 成为多项指标标签 t 中的第 k 个元素项，这 softmax 层的输出作为条件概率 $dk = P(k \mid [c_1, c_2])$ 的表示,因此，交叉熵误差是

$$E_{cE}(p,t;\theta) = -\sum_{k=1}^{K} t_k \log d_k(p;\theta) \quad (5)$$

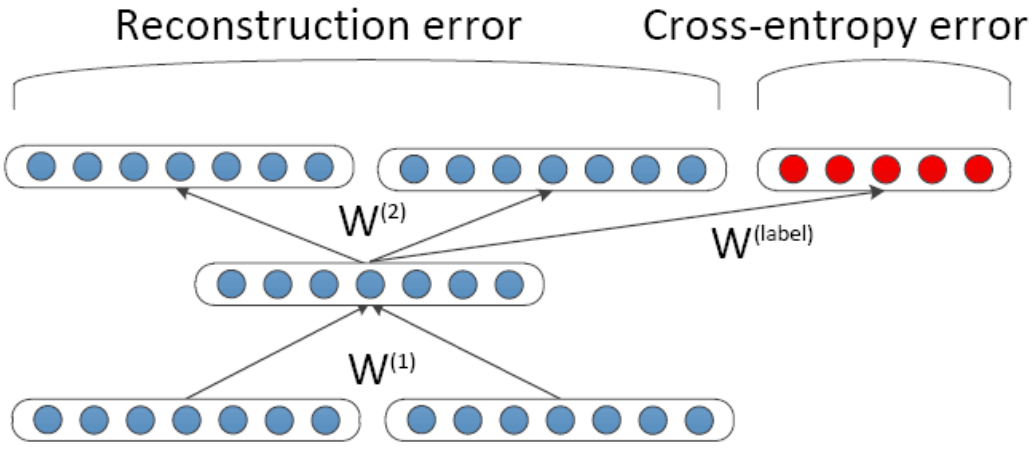


图 12 半监督 RAE 的非终端树节点

半监督的 RAE 最终用下式表示语料库中的每对（句子，标签）：

$$J = \frac{1}{N}\sum_{(x,t)} E(x,t;\theta) + \frac{\lambda}{2}\left\|\theta\right\|^2 \quad (6)$$

每个实体的误差由贪婪 RAE 方法构造的二叉树上的所有节点的误差的总和构成：

$$E\left(x,t;\vartheta\right) = \sum_{s\in T(RAE_\theta(X))} E\left(\left[c_1;c_2\right]_s, p_s, t, \theta\right) \quad (7)$$

每个非终端节点的误差由它的重建误差和交叉熵误差构成：

$$E\left(\left[c_1;c_2\right]_s, p_s, t, \theta\right) = \alpha Erec\left(\left[c_1;c_2\right]_s;\theta\right) + \left(1-\alpha\right)E_{cE}\left(p_s, t;\theta\right) \quad (8)$$

上式中的 α 为超参数，表示节点的重构误差在总误差中所占权重。

使用这个模型时预测句子的情感分布时，利用树的的顶节点的向量表示，并训练简单的逻辑回归分类器。

2) 情感分析过程

a. 构建半监督 RAE 的模型

通过人工标记，得到积极、消极评论各占一半左右的数据集用于模型的构建，将经过预处理和文本矩阵转化的数据集作为输入，通过以下步骤构建半监督 RAE 深度学习模型（本实验中的训练集和测试集是在模型训练过程中按照分别占 60%和 40%随机分配的）。

- 训练模型：训练数据集作为输入，利用 L-BFGs 算法训练模型，实现程序见附件；
- 评价模型：将随机生成的测试集用来测试上一步中构建的半监督 RAE 模型并进行评价在情感分析研究中，常用的评价指标有准确率，召回率，F 值等，本文中采用的是准确率。本次建模的测试结果如图 13 所示，用测试集测试模型，达到了 85.13%的准确率；

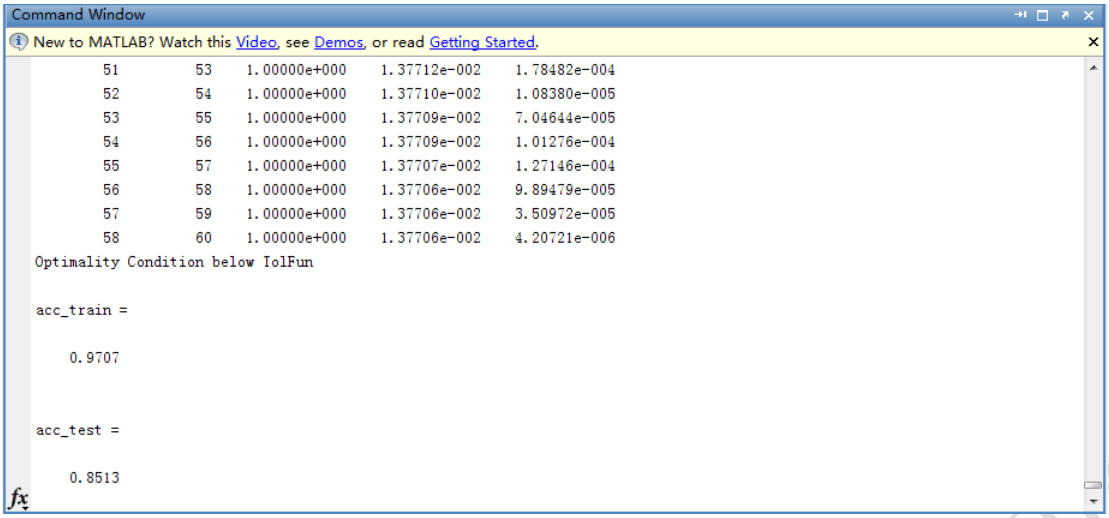


图 13 模型测试结果

◆ 情感分析：利用上一步中构建的半监督 RAE 深度学习模型，分析本次实验的实验数据，分析得到每句评论的情感倾向性，结果如图 14 所示：第 k 列的情感标签表示相应停用词过滤后的评论数据中第 k 行的评论的情感倾向性，0 表示消极，1 表示积极。京东美的 F50-21W6、京东海尔 EC5002-D、京东格兰仕 G50E302T、苏宁美的 F50-21W6 的评论情感分析结果分别存在 mlabel.mat 、hlabel.mat、glabe.mat、Slabel.mat 文件中，数据文件见附件。

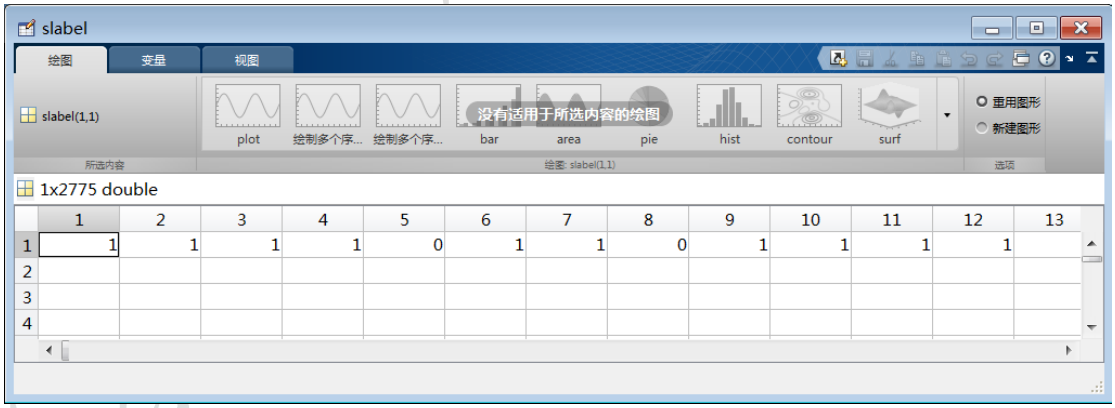


图 14 情感分析结果

步骤五：属性提取并统计

本步骤主要是结合步骤三得到词表和步骤四得到的情感分析结果，进行统计，得到包含某属性的评论数据中积极、消极评论所占的百分比。继而分析用户对产品的某个属性或者电商平台的服务的满意程度。

1) 根据步骤三中生成的词表提取出属性相关词并分类，结果如图 15 所示，每个属性对应的是步骤三中的词表中属性相关词的编号。前面 11 个是电热水

器的属性，后 3 个是电商平台的服务质量的属性。

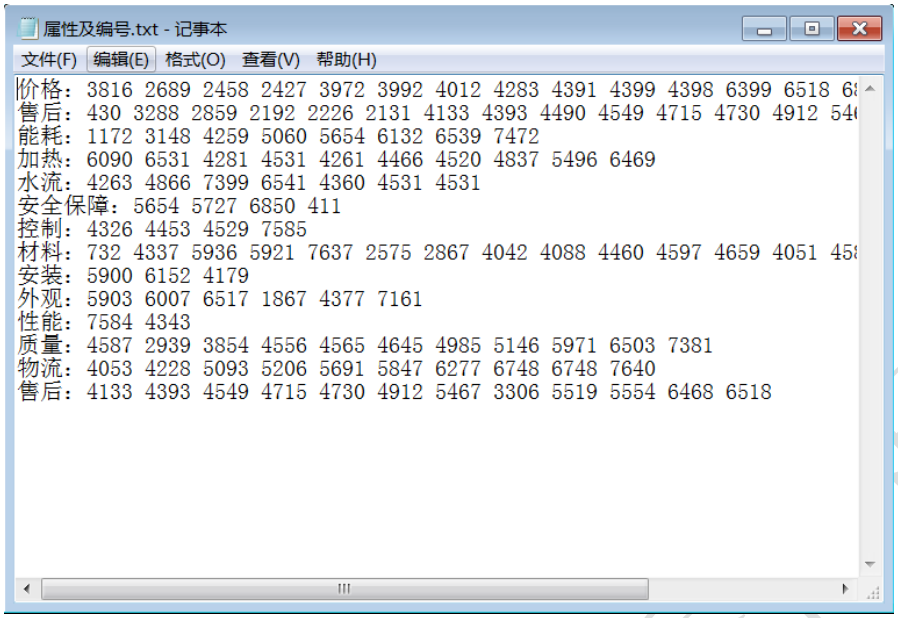


图 15 属性相关词提取结果

2) 利用程序遍历，统计分析得出包含某个属性相关词的评论数据中的积极评论与消极评论的数量，和各自占该商品的与该属性相关的所有评论数量的比重。具体实现程序见附件。

2.3. 结果分析

将上述步骤五得到的结果进行以下几个方面的详细分析：

1) 同一电商平台销售的同一产品的不同属性分析，提炼该商品的竞争优势与劣势，并提出产品改进方案。

a. 京东美的 F50-21W6 的各个属性的积极百分比值比较结果如图 16 所示，该电热水器最大的特点就是能耗较低，而对于美的公司来说，电热水器的加热和控制方面需要进一步的技术改进，以更好的吸引消费者。

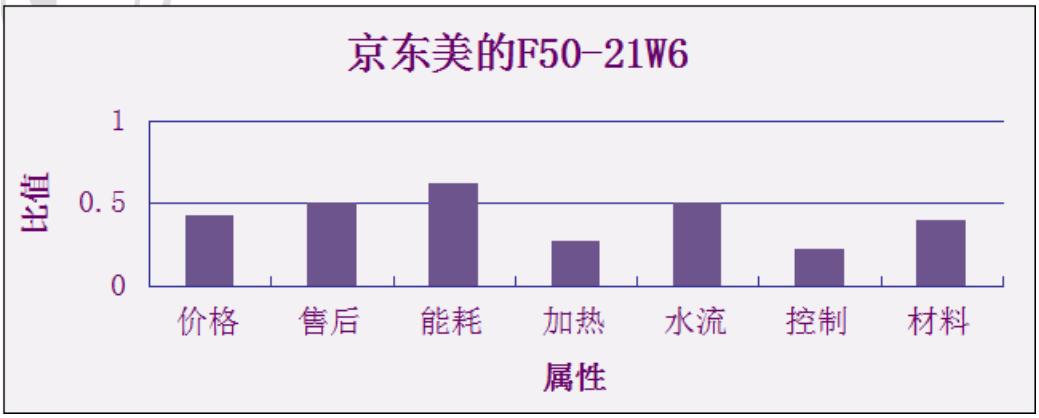


图 16 京东美的 F50-21W6 的各个属性的积极百分比值

b. 京东海尔 EC5002-D 的各个属性的积极百分比值比较结果如图 17 所示，海尔的这款电热水器能耗、价格、加热等方面较有优势，为了提高该产品的销售额，吸引更多消费者，海尔公司应该在产品的外观和使用控制方面进一步的改进。

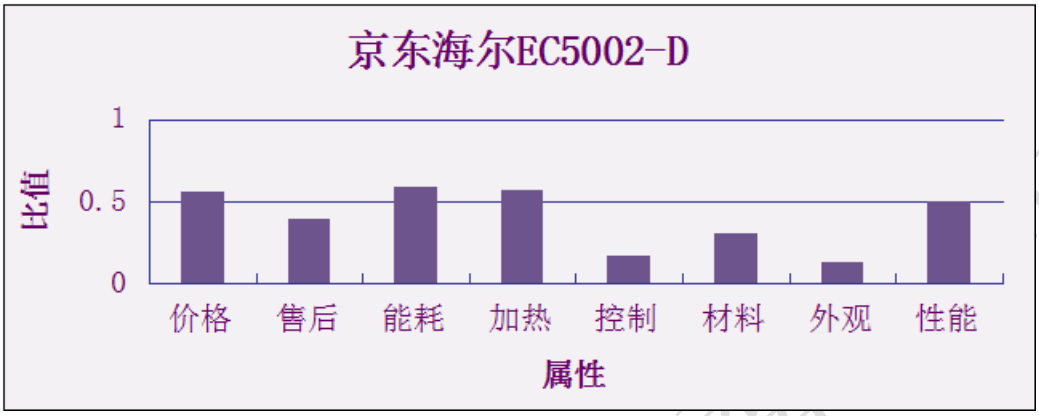


图 17 京东海尔 EC5002-D 的各个属性的积极百分比值

c. 京东格兰仕 G50E302T 的不同属性的积极百分比值比较结果如图 18 所示，格兰仕的该款电热水器最大的特色就是水流方面质量非常好，此外能耗方面也是比较符合用户需求的，该产品在外观和热水器控制方面可以做进一步，提高销售额。

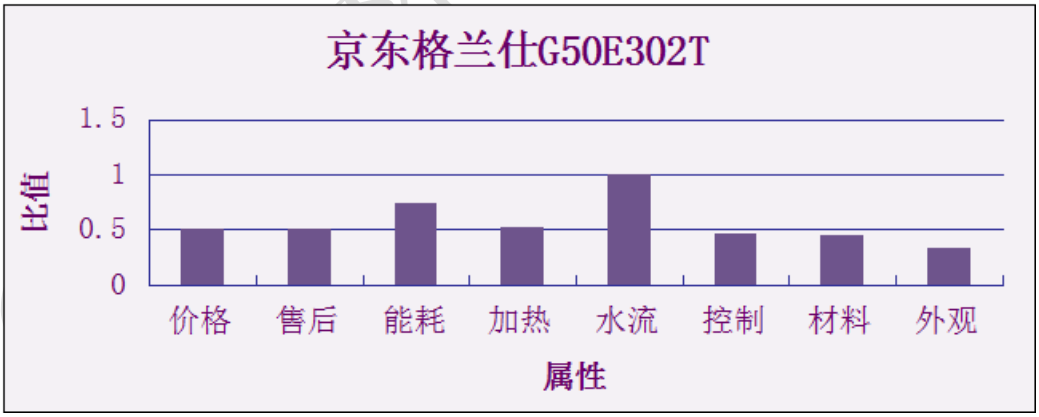


图 18 京东格兰仕 G50E302T 的各个属性的积极百分比值

2) 针对同一个属性，比较不同品牌的电热水器的差别，提炼出各个品牌的差异化卖点。结果如图 19 所示，对于美的该款电热水器来说，竞争优势不是特别大，但是相比海尔的这款产品，在售后、能耗、材料等方面有一定的竞争优势；海尔的这款电热水器价格实惠，用户对它的加热也比较满意，但在其

他方面没有什么优势了；格兰仕的这款电热水器整体上竞争力较强，能耗低较能吸引更多的消费者。

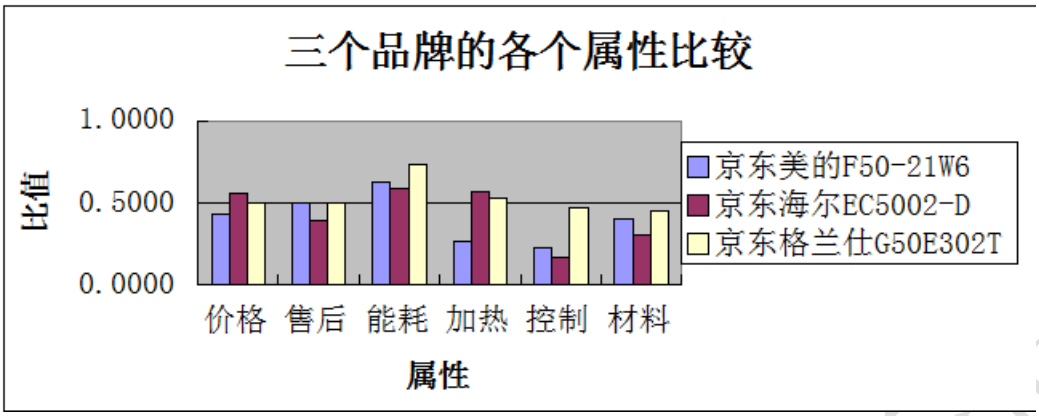


图 19 三个品牌各个属性的比较结果

3) 不同电商平台间的服务质量比较，各个属性相关评论中积极评论百分比结果如图 20 所示，京东在质量、物流、售后、价格方面的服务质量均比苏宁易购的高，说明京东是一个比较受消费者欢迎的购物平台，但在售后和价格方面仍有提升空间；相比之下，苏宁易购没有什么明显竞争优势，该平台可以首先考虑从商品价格和质量上做些改进，以吸引更多消费者（注：由于实验中使用的数据有限，关于电商平台的比较结果可能不具有代表性）。

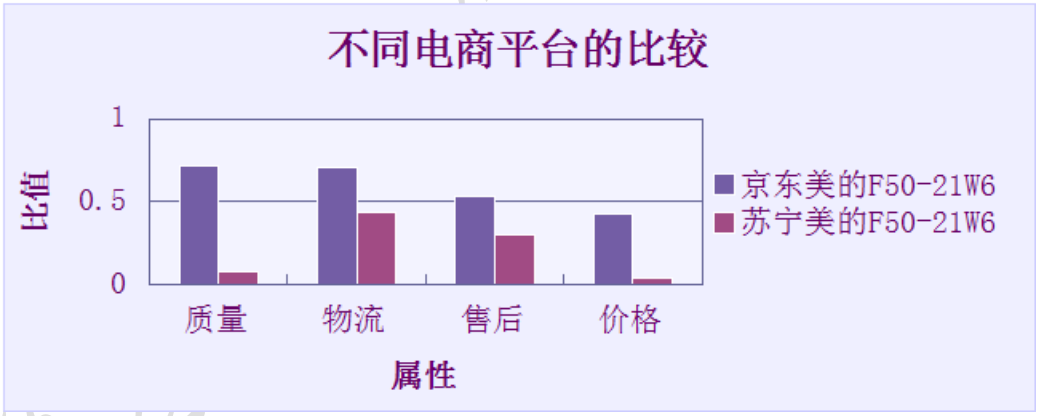


图 20 不同电商平台的比较

3. 结论

总结本次比赛，我们根据网上的电热水器评论数据的特点，利用构建的半监督 RAE 深度学习模型进行情感分析，统计分析出评论数据的情感倾向性以及用户对每个产品或者电商平台的某个属性的满意程度。实现了本次的挖掘目标：提

炼出来不同品牌电热水器的差异化卖点和产品改进方案；得到了两个电商平台的竞争优势与劣势。

本次评论数据挖掘分析的过程中，每一步都通过程序实现，进行了大量的数据挖掘分析工作，实验中的每一步都有理有据，各个步骤之间联系密切，条理清晰且系统地完成了本次数据挖掘分析工作。但是在实验过程中依旧遇到了很多瓶颈问题，例如关于产品或者电商平台的属性分析问题，本次实验中的情感标签分为积极和消极，利用每句评论的情感倾向性，去估测用户对某个属性的满意程度，属于粗粒度的情感分析。在之后的研究学习过程中，我们将继续针对某属性，进行以下两个方面的细粒度情感分析：

1） 将属性的情感标签多级量化进行情感分析。

2） 细粒度分析某属性，因为一个句子中可能对多种属性进行了评论，不同属性的评论可能情感不一致，将一个句子继续细分，分析其中每个属性的情感倾向性。

4. 参考文献

[1]梁军，柴玉梅，原慧斌，等.基于深度学习的微博情感分析[J].中文信息学报，2014，Vol.28 No.5:155-161

[2]朱少杰，基于深度学习的文本情感分类研究.哈尔滨工业大学:硕士学位论文.2014

[3]Socher R, Pennington J, Eric H. H., et al. Semi-Supervised Recursive Autoencoders for Predicting Sentiment Distributions [C]. EMNLP. 2011

[4]张紫琼，叶强，李一军.互联网商品评论情感分析研究综述[J].管理科学学报,2010.Vol .13 No.6:84-96

[5]王继成，潘金贵，张福炎.Web文本挖掘技术研究.计算机研究与发展，2000，Vol.37,No.5

[6]Quoc Le,Tomas Mikolov.Distributed Representations of Sentences and Documents,cs.CL，2014.05

[7]孙莹.基于Web文本挖掘的企业口碑情感分类模型研究.华中师范大学：硕士学位论文.2013.05

[8]王雅思.深度学习中的自编码器的表达能力研究.哈尔滨工业大学:硕士学位论文.2014

[9]徐德.关于互联网文本数据挖掘的一些关键技术研究.电子科技大学：硕士学位论文.2011

[10]祖李军，王卫平.中文网络评论中提取产品特征的研究.计算机系统应用.2014