基于电商产品评论数据的情感分析

张美颀
（沈阳化工大学　辽宁省沈阳市　110142）

摘　要：本文爬取了京东商城中“西门子”品牌的冰箱消费者购买使用后的评论数据，然后通过数据清洗、机器预处理、停用词过滤工作分别处理爬取下来的文本数据[1]。接着通过建立深度神经网络、LDA 主题模型以及语义网络等模型，来判定文本数据所表达的情感倾向以及分析出其深层次的含义[2]。

关键词：数据挖掘；评论情感分析；爬虫技术

互联网行业的高速发展，带来的是人们越来越喜欢在网络上公开分享自己对某一事物的看法或情绪，因此微博、商品评论、电影评论等文本信息大量出现。利用信息，并深入的挖掘和分析，可以为消费者和企业提供相关的参照依据，为提高产品的服务与质量提供了数据依据，也可以为政府提供舆情监测和分析。由于网络上的文本信息量过于庞大，光靠人力是不能够解决问题的，这时就需要运用当今的信息技术。利用数据挖掘、机器学习等技术对网络上大量的文本评论进行处理，然后分析其表达的情感倾向，这就是情感分析。主要的工作有抽取评价对象与短语以及搭配关系。

1 确定评论分析目标

首先通过爬虫技术从京东商城上爬取了“西门子”品牌冰箱的用户评论数据，并根据评论数据集建立基于商品评论数据的情感分析数据挖掘模型，需要进行数据挖掘建模的目标如下：

（1）根据评论数据分析出“西门子”冰箱用户使用后的情感倾向；

（2）根据所爬取的评论数据挖掘出“西门子”冰箱所具有的优缺点；

（3）根据对“西门子”品牌冰箱的评论分析，提炼出其它品牌冰箱的卖点。

2 商品评论分析的方法和过程

本文结合从京东商城上以爬虫技术爬取的“西门子”冰箱的用户评论数据为基础建立模型，分别从数据抽取、数据探索与预处理、建模 & 诊断、结果 & 反馈等几个方面对商品评论数据进行分析，然后通过建立以为语义网络或 LDA 算法为基础的数据挖掘模型[3]，分析出评论数据情感倾向的分类问题以及其深层次的隐藏含义。分析的流程图如图 1 所示。

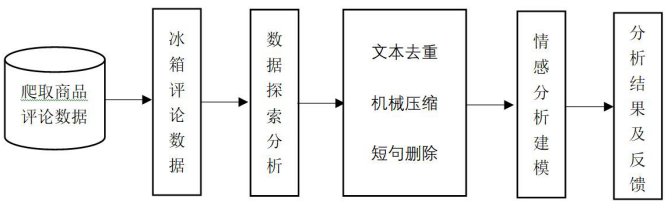

图 1：商品评论数据挖掘分析流程图

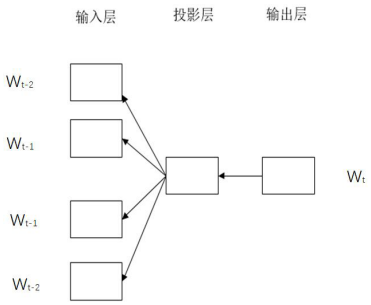

图 2：CBOW 模型

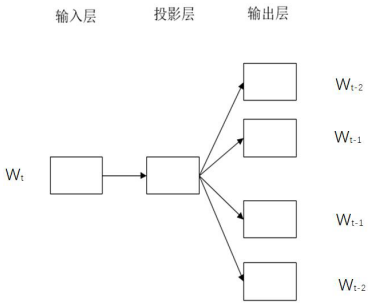

图 3：Skip-gram 模型

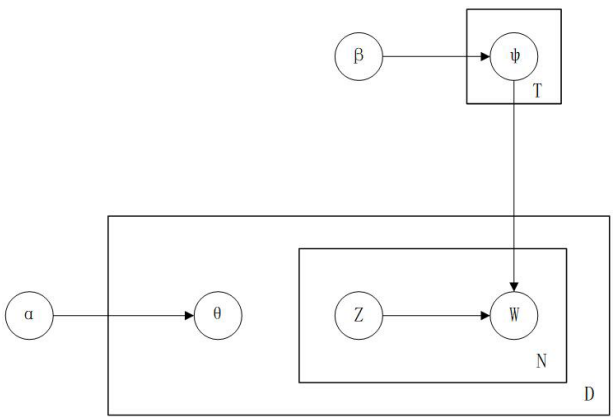

图 4：LDA 生成模型

3 数据预处理

根据图 1 中商品评论数据挖掘分析流程图可知，数据预处理可分为以下 3 个部分，分别为：文本去重复、压缩去词、删除短句。

3.1 商品评论数据文本去重复

此步骤的目的主要是为了去除商品评论数据中的重复部分，这样可以起到去除无用评论和重复评论的目的。一般的文本去重算法的主要思想是利用算法分析文本之间的相似程度，然后根据相似程度的深浅进行文本去重。这类算法包括 Simhash 算法、距离去重等。其中距离去重算法是通过计算两条不同语句间的编辑距离，然后分别对其计算得到的距离阈值进行判断，如果计算得到的编辑距离与阈值之差为负数，那么将进行去重处理。但是当遇到所要表达意思相近的语句时，该语句也可能因为去重算法而被删点，这样就会导致错删的情况出现。为了避免错删，这里我们采用较为简单的去重思路，那就是只对完全重复的语句进行去重。

表 1：数据集描述

数据集统计	数目
评论条数	15137
用户个数	8700
评论 > 五条的用户个数	523

3.2 机械压缩和短句删除

虽然进行了文本去重，但是远未达到数据清洗的目的，因为现实中有一些评论数据中含有一些连续重复的语句，而这些语句并没有什么实际的意义，所以需要对其进行压缩，以去掉一些不必要的表达，东西好用好用好用，只需要压缩为东西好用即可。短句删除的思想来自于，当机械压缩完成去词处理后，应该将过短的句子进行删除。

3.2.1 压缩去重复词

机械压缩可以去除语句中重复的词语，在一般的评论中时常会出现一些连续重复的开头和结尾，因此我们只需要对文本的开头和结尾进行处理即可。例如：“安装费怎么这么这么这么贵”。压缩去重也有其规则，我们可以通过建立两个字符列表来完成，当读取到重复词后，先将重复词放置在第一个列表中，再将下一个重复词放置到第二个列表中，以此类推读取重复字符，当情况不同时触发相应的词语压缩准则，如果再次出现与列表 1 和列表 2 完全相同的次则对其进行压缩。根据以上词语压缩规则，既可以对开头或结尾重复的语句进行压缩处理，这样即可得到较为精炼的语句。

3.2.2 删除较短的评论

对短句进行删除就是通过设置评论字数的最小限制，当评论语句字数小于此最低门限字数限制时即对其进行删除。

4 商品评论情感分析数据挖掘建模

4.1 结合情感倾向建立模型

首先对现有文本进行训练，得到相应的词向量，这样就可以将文本符号数学化 [4]，使文本情感分析的问题转化成了一个深度学习的问题。通常用独热编码、分布式表示来表示词向量，例如 word2vec 模型就是一个分布式表示方法 [5]。Word2vec 的核心思想是通过词的上下文得到词的向量化表示，利用深度学习的思想，将对文本内容的处理，用向量运算表示，即文本语义上的相似度可以用得到的向量空间上的相似度来表示。Word2vec 包含两种框架：

（1）CBOW（通过附近词预测中心词），本文使用 CBOW 框架，结构如图 2 所示。

（2）Skip-gram（通过中心词预测附近词）结构如图 3 所示。

4.2 人工标注数据集

在通过词向量构建得到相应结果后，还需对商品评论文本数据的子集进行人工标注，如果是对商品进行正面积极的评论，那么此评论被标记为 1，反之，若此评论语句是对商品进行反面消极的评论，那么此评论则会被标记为 -1。评论与向量是一一映射的关系，将语句中所有分词的词向量相加之后取平均值，最终得到的词向量的情感倾向值可以判定为评论的情感倾向 [6]。

4.3 基于LDA算法模型对文本主题进行分析

如果从统计学的观点出发，我们将文本中的主题词和特征词进行统计，并对其出现的频率进行量化。在本文中，运用 LDA 算法模型，可以挖掘到更多不同品牌评论中的深层信息。在机器学习和自然语言处理中，LDA 算法模型常被用来统计一些抽象的统计模型。LDA 算法是一种无监督深度学习算法。LDA 模型是一种生成模型，如图 4 所示。

LDA 算法模型也叫层贝叶斯概率模型，将文本分为了文档、语句、词语的层次结构，可以有效的建立起相应的文本概率分析模型。根据 LDA 算法模型 [7] 的分析，能够从文本中挖掘到其潜在的主题，进而能够重点关注文本中的特征词，精确的把控住文本的大概含义。LDA 模型的生成过程是：先确定一篇文档 D，文档和主题、主题和词汇表中的词分别满足两个带有超参数 α 和 β 的多项式分布。θ 代表文档的主题分布，ψ 代表词分布，其过程就是从 θ 中抽取主题，再从其对应的 ψ 中抽取一个词，进行 N 次上述操作后得到文档。这样，就可以把抽象的文本信息转化成能够建立相关数学模型的数字信息 [8]。其概率模型公式如公式（1）所示。

$$p(\theta,z,w|\alpha,\beta)=p(\theta|\alpha)\prod^{N}_{n=1}p(z_n|\theta)p(w_n|z_n,\beta) \quad (1)$$

5 实验与结论

本文采用的数据是从京东商城上爬取下来的用户购买使用后对于西门子冰箱的评论数据，约 15000 条，主要信息包括用户信息，冰箱信息以及文字信息。其中用户信息包括用户的名称、等级；冰箱信息包括外观外形、制冷效果、保鲜效果、容量大小、噪音大小等；文字信息包括评论者、使用反馈、物品评分、纯文字等，表 1 是数据的描述。

在数据挖掘、文本聚类等方面，LDA 算法模型被广泛的应用 [9]。相比于其它的文本情感分析神经网络模型，它引入了狄利克莱函数的先验概率信息，因此在文本情感分析过程中，该模型有较强的泛化能力，很少产生过拟合。并且该方法是一种无监督的深度学习方法，在只要提供预料数据集的情况下，就可以自动分析并训练出情感文本 [10] 的各种情感倾向概率。此算法对于电商评论的情感分析能起到较好的作用。

参考文献

[1] 李旭，于卫红．基于情感分析和关系网络的影视产品评论数据文本挖掘研究 [J]．情报探索，2018(4)．
[2] 曹鑫．商品文本评论数据情感分析研究 [J]．字化数用户，2017(39)．
[3] 康晓东．统计类数据挖掘和知识类数据挖掘 [M]．北京：机械工业出版社，2004:187．
[4] 陈涛．基于分布式表示学习的文本情感分析 [D]．哈尔滨：哈尔滨工业大学，2017．
[5] 方振宇．基于词向量的微博用户抑郁预测方法研究 [D]．合肥：合肥工业大学，2017．
[6] 刘姗，胡勇．中文网络话题评论文本语义倾向分析 [J]．信息安全与通信保密，2012(6)．
[7] 邹晓辉．LDA 主题模型在文本聚类中的应用 [J]．数字技术与应用，2017(12)．
[8] 崔磊，周明．统计机器翻译领域自适应综述 [J]．智能计算机与应用，2014(6)．
[9] 汪丹丹．中文文本聚类算法研究 [D]．苏州：苏州大学，2016．
[10] 高宇．基于跨媒体的社交电商用户情感分析 [D]．哈尔滨：哈尔滨商业大学，2018．

作者简介

张美颀（1995-），女，研究生在读。研究方向为自然语言处理。