

# Application Research of Text Mining in Commodity Reviews

## —Taking Tobacco Reviews as an Example

Chunguang Jia

School of Statistics and Mathematics, Yunnan University of Finance and Economics, Kunming Yunnan  
Email: 1106278959@qq.com

Received: Dec. 6<sup>th</sup>, 2018; accepted: Dec. 19<sup>th</sup>, 2018; published: Dec. 26<sup>th</sup>, 2018

---

### Abstract

With the rapid development and popularization of Internet technology, the Internet provides a lot of places for users to comment on products. These comments directly reflect the customer's emotional attitude towards the function or performance of the product. Therefore, text mining of product reviews is of great significance. However, the amount of online commentary data is huge, mostly semi-structured and unstructured data, and there are many useless comments. How to quickly obtain commodity review corpus and select which method to analyze becomes a key issue for research. First of all, this paper uses Python to obtain tobacco commentary corpus through crawlers, and performs data preprocessing operations such as simplification and corpus transformation, typos replacement, useless comment culling, etc., and then based on the preliminary categorization of the corpus into positive and negative emotions. The emotional score of tobacco is calculated based on the sentiment dictionary, the degree adverb dictionary, and the negative word dictionary. The results show that the domestic emotional scores on this commodity are still relatively high, and the scores of the provinces along the Yangtze River are slightly higher than other regions.

### Keywords

Web Crawler, Text Mining, Sentiment Analysis, Tobacco

---

# 文本挖掘在商品评论中的应用研究

## ——以烟草评论为例

贾春光

云南财经大学, 统计与数学学院, 云南 昆明  
Email: 1106278959@qq.com

收稿日期：2018年12月6日；录用日期：2018年12月19日；发布日期：2018年12月26日

## 摘要

随着互联网技术的飞速发展与普及，网络上提供了很多用户对商品评论的地方，这些评论信息直接体现了客户对商品功能或性能方面的情感态度，因此对商品评论进行文本挖掘具有重大意义。然而网络评论数据量巨大，多半为半结构化、非结构化数据，且其中的无用评论较多，如何快速获取商品评论语料以及选取何种方式分析成为研究的关键问题。首先，本文利用Python通过爬虫获取烟草的评论语料，并对语料进行简繁转化、错别字替换、无用评论剔除等数据预处理操作，接下来在把评论语料初步分为正面情感和反面情感的基础上，基于情感词典、程度副词词典、否定词词典计算消费者对烟草的情感评分。结果表明：国内对本商品的情感评分还是比较高的，且长江沿岸省份的评分稍高于其他地区。

## 关键词

网络爬虫，文本挖掘，情感分析，烟草

Copyright © 2018 by author and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

## 1. 引言

### 1.1. 研究背景与意义

#### 1.1.1. 研究背景

随着社会的发展和时代的进步，信息社会的发展速度已经超出了绝大多数人的想象，与此同时，互联网容量也达到了一个空前的规模，互联网已经成为人们日常生活中必不可少的一部分[1]。

据中国互联网信息中心(China Internet Network Information Center, 简称CNNIC) [2]于2018年1月31号发布的第41次《中国互联网发展状况统计报告》显示，截止2017年12月，中国互联网的用户约为7.72亿，互联网普及率更是高达55.8%，中国的互联网行业开始向标准化、商业化发展，而移动互联网的快速发展，进一步促进不同场景的设备智能化、消费模式的共享化和评论的多元化。

在购物方面，消费者的计划或者决策行为受其他消费者评价的影响较为明显。以往的实证研究表明，口碑传播对消费者行为起到了非常重要的作用。在信息技术较为成熟的今天，在线口碑具有快捷和方便的特征，能够摆脱地域限制实现多对多的信息传递，因此，更具影响力[3]。

本次用到的是烟悦网(国内最大的烟民交流网站)上的相关评论，首先，可以从天猫、京东中得到启示，把评论文本简单的归为好评、中评、差评来展示。客户可以通过评论的总量、好评和差评的情况来大致了解商品的情况，而对于生产厂商来说，烟悦网上的评论可以大致反映不同商品在客户心中的地位，生产商可以通过分析已购买商品的客户评论，了解商品的正面评价或者负面评价以及商品新的需求点。挖掘出客户对商品的情感倾向性之后，制造商就可以从更贴近市场需求的角度来开发所生产的商品，制造出更具有竞争力的商品，继而提高在市场上的占有率。

#### 1.1.2. 研究意义

随着电子商务模式的日益成熟和网络购物的流行，互联网上保存了大量客户参与的、对于商品研究

有较大价值的评论性文本信息。这些评论性文本直接表达了客户对商品所持的褒贬态度，但这些有用的信息常常被大量的无关评论信息掩埋，故本文在剔除无关的文本信息以及无用的评论内容的基础上，基于情感词典对评论内容进行情感极性分析，分析不同地域对商品的认知情况。

## 1.2. 研究现状

文本情感是指对用户发表的主观性文本信息进行观点、喜好以及情感倾向性分析、监测和挖掘，通常涉及自然语言处理、计算机语言学、信息检索、统计学、人工智能等多个学科研究领域。目前，文本情感分析领域主要的研究工作包括情感信息的获取、分类、检索和归纳。依据研究方法上的差异，可以把其分为如下两类：基于规则的情感分析和基于机器学习的情感分析。

### 1.2.1. 基于规则的情感分析

在国外的研究中，Hatzivassiloglou (1997) [4]等对大规模语料数据集的研究发现英文中形容词的语义情感倾向性往往受到连词的影响，并利用该方法尝试对英文形容词性单词进行情感极性判别。Turney (2002) [5]通过筛选种子词语集合，对评论性文本分三个步骤抽取特征并计算了所获得的特征短语语义倾向的平均值，该方法在汽车类文本与电影类文本的准确率分别是 84%、65.83%。

在国内的研究中，朱嫣岚(2006) [6]等在 HotNet 提供的语义相似度以及语义相关场概念的基础上，计算被测词与评价种子词的相似性差异来确定被测词的语义倾向。林鸿飞(2007) [7]等经过对文本的词汇与结构分析，获得了 9 个对文本情感影响较大的语义特征，接下来利用人工与自动获取相结合的方法进行文本情感分析的研究。何婷婷(2010) [8]等提出了一种以语义理解为基础的文本情感分类方法，继而在情感词的判别中引入情感义原，并给予情感词概念情感语义，同时重新定义文本的情感相似度，获得情感词的情感语义值，此外，还进一步分析了语义层副词的出现对文本情感极性的影响，该方法大大提升了文本情感极性的识别能力。

### 1.2.2. 基于机器学习的情感分析

目前，基于机器学习的情感分析研究中，比较成熟的研究算法有简单的朴素贝叶斯分类算法、决策树算法、KNN 算法、SVM 算法、神经网络算法等。

在国外的研究中，Pang (2005) [9]等在对电影评论的研究中，其利用机器学习中的朴素贝叶斯、最大熵、支持向量机方法对评论语料进行正面和反面的情感分类，与手工分类相比，研究表明选择 SVM 的分类效果最好，其精确度达到了 84%。

中文方面，唐慧丰(2007) [10]等通过用形容词、副词、名词和动词做不同的文本表示特征，以文档频率、信息增益、CHI 统计量和互信息作为特征的选择方法，采用贝叶斯分类、KNN、类中心向量法和 SVM 作为分类器进行试验比较，其结果表明：在有足够大的训练集与选择合适文本特征的前提下，利用信息增益特征选择、支持向量机和 n-Gram 特征表示的实证中，支持向量机取得较好的情感分类结果。黄永文(2009) [11]等基于商品的标准文档研究分析来获取商品的特征属性及其关系，然后在 Bootstarp 的弱监督机器学习基础上获取商品的描述性特征和规格特征的层次关系，接下来利用提供的商品特征属性作为种子词，从文本中自动获取商品的特征情感属性，继而利用抽取的模式获新的商品特征和情感词的组合，可以把该方法看成半自动的情感分类方法。

## 2. 相关技术与理论阐述

### 2.1. 网络爬虫技术概述

#### 2.1.1. 网络爬虫的概念

网络爬虫(Web Crawler)，又称网络机器人、网络蜘蛛，是一个能够自动从网页上获取内容的程序或

脚本。

网络爬虫是搜索技术的核心模块，被广泛应用于检索互联网上各种网页信息，以帮助用户获取互联网相应内容。在实际应用中，一般网络爬虫由控制器、解析器和资源库三个部分组成。控制器担当的职责是给多线程中的各个爬行线程分配工作任务；解析器最主要的任务是下载网络页面，而且会对页面的内容进行处理，通常情况下是把一些 JS 脚本标签、CSS 代码内容、空格字符、HTML 标签等内容处理掉，爬虫的基本工作是由解析器完成；资源库是用来存放下载的网页资源。

由于论坛、微博博文和购物网站上的商品信息、用户评论等 URL 具有一定的规律，通常可以使用软件程序(例如 Python、R、Java 等)进行网页内容的获取和解析，然后使用特定的方法(例如正则表达式等)从爬取的页面中批量获取自己需要的内容，最后将抽取的内容保存到存储文档或者数据库(例如 MySQL、MongoDB 等)中[12]。

### 2.1.2. 网络爬虫的分类

网络爬虫根据爬取方式和实现技术[13]的不同，通常可以分为以下几个类型：通用网络爬虫(Universal Web Crawler)、主题网络爬虫(Theme Web Crawler)、增量网络爬虫(Incremental Web Crawler)和深层网络爬虫(Deep Web Crawler)。

#### 1) 通用网络爬虫

常常利用一些种子 URL 作为基础，通过爬虫器自动发现新的 URL 从而扩充到整个 Web。该爬虫器的目标是采集整个 Web，因此对硬件的要求比较高，需要足够的内存以及存储数据的硬盘空间，而对采集页面的顺序要求相对较低[14]。

#### 2) 主题网络爬虫

是指有选择性的获取与先前设定的主题有关系的爬虫框架。由于其只抓取与主题相关的内容，从而大大的节约了硬件与网络资源的占用，且获取内容的速度非常快，其可以较好的满足对某一领域有特殊需求的人群批量获取内容。主题爬虫按链接的访问方式不同可以分为：基于网络内容评价的爬行方法、基于链接结构评价的爬行方法、基于增强学习的爬行方法以及基于语境图的爬行方法。但如何保证数据的精确性和完整性，都是有待解决的问题。

#### 3) 增量网络爬虫

增量式网络爬虫是指对已经访问过的内容以逐渐递增的更新方式获取新产生的或者发生变化的网页。该类型爬虫的优点是仅仅只在需要时才会爬行新产生或发生变化的网页，并不会重复抓取没有发生任何变化的网页，而且可以非常有效的缩减爬行时间和存储空间上的资源消耗，但不足之处是爬虫程序算法的复杂度和实现难度大大增加了。确保本地存储页面中的内容最新以及提高本地页面中内容的质量是增量式爬虫两个非常重要的目标。

#### 4) 深层网络爬虫

网络页面按搜索引擎发现的难易水平能够分为表层网页(Surface Web)和深层网页(Deep Web)。表层网页是指利用常用的搜索引擎通过超链接非常容易访问到，主要是由静态页面构建的网络页面。深层网页是指在万维网中其网络页面内容常常不可以利用静态网址直接访问、隐藏在索引表单后的，只有访问者上传一些关键词后才可以获取内容的网页，如暗网。深层网络爬虫常用的表单更新方式有两种类型，即以领域知识为基础的表单更新和以网页结构分析为基础的表单更新。

## 2.2. 中文分词技术概述

### 2.2.1. 中文分词简介

在英语中，单词本身就是“词”的表达，一篇文章就是“单词”加分隔符(空格)来表示的，这和汉语

所表达的“词”的概念有所不同。在汉语中，词是以字为基本单位的，但是一篇文章的语义表达却仍然是以词来划分的。因此，在处理中文文本时，需要进行分词处理，将句子转化为词的表示，这个切词处理过程就是中文分词(Chinese Word Segmentation) [15]。

### 2.2.2. 中文分词的主要流派

自中文分词被提出以来，历经将近 30 年的探索，提出了很多方法，可主要归纳为“规则分词”、“统计分词”和“混合分词”这三个主要流派。

#### 1) 规则分词

基于规则的分词是一种机械的分词方法，主要通过维护词典来进行，在切分语句时，将语句的每个字符串与词表中的词进行逐一匹配，找不到则切分，否则不予切分。

根据匹配切分的方式，分为正向最大匹配法(Maximum Match Method)、逆向最大匹配法(Reverse Maximum Match Method)以及双向最大匹配法(Two-Way Maximum Matching Method)三种方法。

正向最大匹配法其算法描述如下：

① 从左到右取待切分汉语句的  $m$  个字符作为匹配字段， $m$  为机器词典中最长词条的字符数。

② 查找机器词典并进行匹配，若匹配成功，则将这个匹配字段作为一个词切分出来，若匹配不成功，则将这个匹配字段的最后一个字去掉，剩下的字符串作为新的匹配字段，进行再次匹配，重复以上过程，直到切分出所有词为止。

逆向最大匹配法的基本原理与正向最大匹配法的相同，不同的就是分词切分的方向与其相反。

双向最大匹配法是将正向最大匹配法得到的分词结果和逆向最大匹配法得到的分词结果进行比较，然后按照最大匹配原则，选取词数切分最少的作为结果。

#### 2) 统计分词

随着大规模语料库的建立，统计机器学习方法的研究和发展，基于统计的中文分词算法渐渐成为主流。

基于统计的分词，一般要做如下两步操作：

① 建立统计语言模型。

② 对句子进行单词划分，然后对划分结果进行概率计算，获得概率最大的分词方式，常用的方法如隐含马尔科夫(HMM)、条件随机场(GRF)等。

#### 3) 混合分词

混合分词，顾名思义，就是规则分词和统计分词相结合的一种分词方法。在实际工程中，多是基于一种分词算法，然后用其他分词算法加以辅助，最常用的方式就是先基于词典的方式进行分词，然后用其他分词算法进行辅助。如此，能在保证词典分词准确率的基础上，对未登录词和歧义词有较好的识别。

### 2.2.3. 中文分词工具

本文主要的介绍的分词工具是 Jieba，Jieba 官方提供了 Python、C++、R、iOS 等多平台多语言支持。Jieba 分词结合了基于规则和基于统计两类方法，它提供了三种分词模式：

1) 精确模式：试图将句子最精确地切开，适合文本分析。

2) 全模式：把句子中所有可以成词的语句都扫描出来，速度非常快，但不能解决歧义。

3) 搜索引擎模式：在精确模式的基础上，对长词再次切分，提高召回率，适合用于搜索引擎分词。

## 2.3. 情感分析技术概述

情感分析(Sentiment Analysis)是指对服务、事件、舆论、产品等这些情感色彩比较浓重的主观性文本

进行分析、处理、归纳以及推理的过程。根据研究目的的不同，通常把情感分析分为情感极性分析、情感强度分析以及主观分析等。本文研究的主要内容是分析客户对卷烟的情感倾向，即对他们的情感强度进行评分。

在情感分析的研究中经常使用自然语言处理、文本挖掘、机器学习以及统计学等多个不同学科的知识。目前，在国内，比较权威的语料库只有谭松波整理的酒店评论语料库和中文倾向性分析评测会议(COAE)公开的实验语料，其它评论性文本情感分析语料库较少。语料库大都是从网上获取的，本文的白沙评论语料库就是从烟悦网上爬虫获取的。因网络用语的多用性，通常需要对语料库进行错别字的修改、分词、去除停用词等数据预处理过程。因此，在情感分析的研究过程中形成了一套比较完整的流程：即语料库的获取、数据预处理、情感极性分析。

### 3. 白沙评论预处理和情感词典的构建

#### 3.1. 语料的获取

##### 3.1.1. 研究的对象

本文的目的是从不同地域、不同方面对白沙(和天下)进行情感分析，具有较强的时效性。首先从烟悦网(<http://www.yanyue.cn/>)上爬取相关的评论，而烟悦网是国内最大的烟民交流网站。评论抓取时间为自商品上架到2018年11月中旬，相关评论的详细网址如图1。



Figure 1. Homepage of Yanyue.com

图 1. 烟悦网的首页

##### 3.1.2. 数据获取

本文爬取数据所用的软件为 Python，Python 相比较其他编程语言而言，第一是它抓取网页文档接口更简洁；第二是它本身提供了很多简洁的文档处理功能[16]。

爬虫主要分为三步：抓取页面、分析页面和存储数据。首先是抓取页面，主要使用的是 Requests 库，在抓取网页代码后，下一步就是从网页中提取信息，这里主要使用 BeautifulSoup 库和正则表达式相结合的解析方法，这样既快速又准确的提取出有用信息。最后将这些信息存储到文档中。

最终抓取出来的部分数据如表 1。

#### 3.2. 语料的预处理

##### 3.2.1. 评论的预规范

烟悦网站上有很多评论是具有随意性的，它们大多数为非规范语言、句法结构较为混乱，且文本中

又含有错别字、繁体字、重复评论、无用评论、广告等价值信息较低的评论。因此，不能对获取的文本语料直接进行分词。首先要做的是修改语料库中的错别字、去噪声、繁化简、剔除重复和无用评论。

目前，对文本中出现的错别字只能人工检查和修改。接下来剔除一些重复和无用评论。如“极品极品”、“此评论涉及广告、转让、求购等，已被删除”等。

**Table 1.** Some sample data

**表 1.** 部分评论数据样例

评论	评论者所在的省份
和天下，非常纯。烟吐出来比一般烟更多。空心过滤嘴。软金砂口感的加强版。女人抽还行！	四川省
感觉一般，外观不错	湖南省
包装精美，初尝味道柔和醇甜相当有档次，甚至偶尔感觉接近蓝熊的香醇口感，不错！	广东省
烟还可以，性价比不高。	广东省
这个烟真心不错。。烟嘴空芯。。。我喜欢。。现在也是口粮之一。推荐给大家。。这款烟不愧为白沙之最。很香。。。现在一直抽。	河北省
味道，包装一般，不过仁者见仁智者见智，看个人口味，应该当礼品烟	甘肃省
不好抽，就抽一个逼格，真正的还是建议南京九五至尊或者黄鹤楼	江西省
烟气非常细腻口感甘甜绝对的好烟说啦嗓子的你是买到假烟了	北京市
味道很香，感冒抽都不难受，外观低调奢华，刚看到以为就 40 多的样子，结果是百元档的（’	浙江省
湖南长沙人，真心觉得这烟难抽，与同价位的其他烟相比，性价比不是一般的低，与它一样的还有硬蓝芙	湖南省

### 3.2.2. 文本分词

在汉语中，单单的一个词并不足以表达一个意思，且它们之间没有明显的分隔符，因此，需要对文本语料进行分词处理。分词的结果的对后续的情感分析有着不可忽视的影响。如果分词不佳，即使后面的算法再完美也无法得到好的效果。

本文采用 Python 软件中的分词工具模块“Jieba”进行分词。由于评论表达的多元化，且烟草专有名词的存在，在分词的过程中需要自定义用户词典。搜狗拼音输入法作为荣获多个国内软件大奖的电脑中文输入法生产厂商，其中会有一些关于烟草专业名词的收录，本文根据搜狗拼音输入法的 230 种香烟词库，并辅以自己建立的词典，作为分词的用户字典进行分词。

对文本分词时会出现一些频率极高或者极低的介词、代词、虚词等词语或者特殊符号，它们与后续的情感分析没有什么关系。如“你”、“我”、“他”、“的”、“……”等。为了降低计算机的运算复杂度，提高计算机的运行效率。因此，文本分词时需要建立停用词表直接对这些词语进行删除。

在文本分词中，会出现一些英文单词。比如“FB”、“VERYGOOD”等。遇到这种情况可以特殊处理。“FB”在烟草行业中一般有“要面子”的意思，“FB”烟是高价烟、腐败烟的简称，而“VERYGOOD”可以翻译成“非常好”这样的中文词汇进行分词。

分词效果如表 2。

### 3.3. 基于情感词典的情感分析

#### 情感词典的构建

情感词是指在文本语料中表示情感色彩的名词、形容词、副词、动词、感叹符号和常用的习惯性表达或者短语等。它常常被用来表达文本的情感倾向，文本情感分析最基础的就是构建一个适用性程度高

的情感词典，因此，情感词的研究和分析就成了研究文本情感分析的最重要的环节[17]。

**Table 2.** Example of word segmentation effect

**表 2.** 分词效果样例

感觉 一般 外观 不错
包装 精美 初尝 味道 柔和 醇甜 有档次 感觉 接近 蓝熊 香醇 口感 不错
还可以 性价比 不高
这个烟 真心 不错 烟嘴 空芯 喜欢 口粮 推荐 这款烟 不愧为 白沙之最 很香 一直抽
味道 包装 一般 仁者见仁 智者见智 口味 当当 礼品烟
不好抽 抽 逼格 建议 南京九五至尊 黄鹤楼
价格 太高 味道 一般 性价比 不高
醇香 浓烈 一款 好烟 满足感 很强
这烟 不错 很纯 湖南 很多 老板 抽 湖南 还 卖得 贵点 120 一包
不得不说 价位 这烟 在我心中 仅次 大重九 唯独 一点 涩 算是 这烟 特色 稍稍 探 舌尖 抽 感觉 烟气温热 香甜

通过阅读国内外的文献了解到，文本情感倾向分析研究领域还没有一个可以通用的情感词典，在国外，由 Princeton 大学的心理学家，语言学家和计算机工程师联合设计了一种基于认知语言学的英语词典 WordNet，它不是光把单词以字母顺序排列，而且按照单词的意义组成一个“单词的网络”。哈佛大学在 1966 年整理编著了 GI (General Inquirer)词典，并在其中列出了每个词的情感属性，是目前英文文本分析中经常选用的资料之一。尽管国内在文本挖掘领域的研究起步比较晚，有关中文的情感词典比较稀少，虽是如此，但还是有专家和学者已经整理出很多实用性较强且权威的情感词典。下表列出了国内主要的情感词典资源，如表 3。

**Table 3.** Main emotional dictionary in China

**表 3.** 国内主要的情感字典

词典类别	简介
NTUSD 情感词典	台湾大学整理编著的中文情感词典 NTUSD，它有两个版本，一个是简体中文版，另一个是繁体中文版。每个版本都包括了 2810 个正面情感词汇和 8276 个负面情感词汇。
知网 HowNet	中科院的董振东教授耗时多年建立的中国第一个电子知识系统。它包含了 6 个中文情感分析词集和 6 个英文情感分析词集，并将每个词集划分为 6 类：正面评价词语、正面情感词语、负面评价词语、负面情感词语、主张词语、程度级别词语。
《哈工大信息检索研究室同义词词林扩展版》	本词表是《同义词词林》的扩充，是由哈尔滨工业大学整理编著的，一共包含了 77343 条词语。在《同义词词林》的三层分类系统的基础上，增加了两层编码，一共具有五层结构，并提供五级编码。目前已经推出了 1.0 版本，可以满足很多领域研究的应用。
《情感词汇本体》	该情感词汇本体由大连理工大学信息检索研究室独立整理标注完成。

情感词的收集和研究是一个循序渐进的过程。在目前的文本情感研究中，一般把情感词分为表示褒义的正面情感词以及贬义的负面情感词。因此，本文根据这两方面并结合情感词典对语料中的评论进行评分。

在中英文中，都有用某种修饰手法来表达自己的情感程度或情感极性反转的习惯，例如英语中的比较级、最高级、否定式等修辞形式，中文中也常用“非常”、“特别”、“相当”、“不”、“没有”等程度副词或者否定词来修饰。

程度副词是对一个形容词或者副词在程度上加以限定或修饰的副词，一般位置在被修饰的形容词或者副词之前。在文本评论中，评论者为了强调自己对商品的喜恶程度会大量使用副词，这种程度副词的表示会对判断用户的情感有着重要的作用。如“朋友拿来抽过。很不错。特别喜欢。但是特别贵哦。”在这句话中，有两个程度副词“特别”分别修饰了“喜欢”和“贵”，“很”修饰了“不错”，这些修饰词都表达了一种更加强烈的情感倾向。如果去掉了这些程度词的话，就没有那么明显的情感态度。因此，程度副词在中文文本分类中不可或缺。

本文根据知网 HowNet 的 219 个程度级别词语并查找相关资料进行程度副词词典的构建，用于计算情感得分。蔺璜等人提出可以把程度副词设定为极量、高量、中量、低量四个不同的等级，并人为的设定不同程度副词的权值。根据这些程度副词代表的感情强度赋予它们不同的值，从强到弱分别取 2.0、1.7、1.2、0.6，具体情况如表 4。

**Table 4.** Degree adverbs dictionary

**表 4.** 程度副词词典

程度	权值	个数	词汇
极量	2	72	百分之百、倍加、备至、不得了、多多、不可开交、不亦乐乎、不堪、不折不扣、胸、充分、到头、地地道道、彻头彻尾、非常、极、极为、逾常、截然、尽、惊人地绝、绝顶、绝对、绝对化、刻骨、酷、满、满贯、满心、莫大、奇、入骨、甚为等。
高量	1.7	69	不过、不胜、大为、惨、沉、尤其、沉沉、特、出奇、多、多么、分外、格外、够戗、好、好不、何等、很、够瞧的、很是、坏、可、老大、良、超微结构、太甚、特别、尤、多加、尤为、尤以、远、着实、曷、侈、不为过、超、超额、超外差、超物质、出头、多、浮、过、过度、过分、过火、过劲、不少、过了头、过猛等。
中量	1.2	37	如斯、足、益、这般、足足、尤甚、逾、愈、更加、愈……愈、愈发、愈来愈、远远、越……越、越发、越加、越来越、越是、这样、益发、愈加、大不了、愈益、多、更、更进一步、更为等。
低量	0.6	41	半点、一点儿、不大、不甚、一些、不怎么、相当、聊、轻度、些小、弱、丝毫、微、相对、稍为、稍许、挺、没怎么、未免、些、些微、一点、有点、不丁点儿、有些等。

除了程度副词，否定词也是副词的一种，一般为表示否定意义的词语，根据中文表达常识可知，一般情况下否定词修饰的情感词其极性会发生反转。如“味道很香，感冒抽都不难受，外观低调奢华，刚看到以为就 40 多的样子，结果是百元档的”这句话，“难受”在一般文本中表达的情感是负面的，但在其前面加上了“不”就变成了正面情感。但在中文语句中含有多重否定的句法，当否定词在句子中出现的次数是奇数时，表示否定意思；当否定词在句子中出现的次数是偶数时，表示肯定意思。结合本文的语料库和中文表达习惯，本文收集了 71 个否定词，其权值设定为-1，具体情况如表 5。

**Table 5.** Negative word dictionary

**表 5.** 否定词词典

权值	个数	词汇
-1	71	不大、否、不丁点儿、未尝、未曾、从未有过、不甚、不怎么、聊、没怎么、不可以、无须、怎么不、几乎不、非、从来不、从不、放弃、不用、不曾、不该、毫无、不必、不会、不好、不能、很少、极少、没有、不是、难以、放下、扼杀、终止、停止、反对、缺乏、缺少、不、不要、甬、勿、别、未、反、没、木有、无、请勿、并非、决不、永不、毋、莫、从未、尚未等。

## 4. 情感分析

### 4.1. 情感评分的计算

本文根据 Jieba 分词结果、程度副词词典、否定词词典、BosonNLP 情感词典为基础，制定一套打分

规则，其中 BosonNLP 情感词典是基于微博、论坛、新闻等数据来源构建的情感词典。

从第一条评论的第一个词开始，如果这个词出现在 BosonNLP 情感词典的评分中，就会有如下判断，前后词语是否有程度副词，如果有，就乘以程度副词的权值，如果还有否定词，就在前面乘以权值-1，直到这个向量词组中没有情感词。具体流程图如图 2。

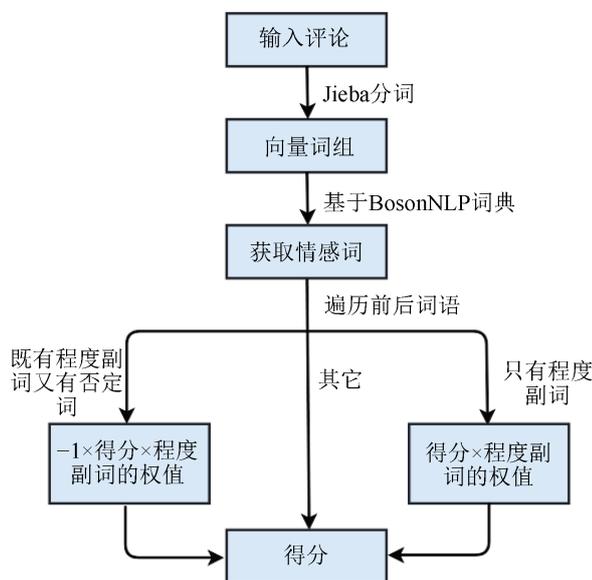


Figure 2. Flow chart for calculating emotional scores

图 2. 计算情感得分流程图

根据得到的分数进行划分，按照得分大于 0、等于 0、小于 0 分为积极情感、中性情感、消极情感三类，分别取值 1、0、-1。部分结果如表 6。

Table 6. Samples of sentiment classification results

表 6. 情感分类结果样例

评论	评论者所在的省份	评分	情感	取值
绵软有力，湘烟代表，作为一名山东汉子，非常喜欢这款烟，外形精致，有一种古典雅韵，适合中年男士，非常不错	山东省	24.54802	积极	1
我们这刚有货，有档次，就是觉得烟嘴太长了	贵州省	0.278466	积极	1
时尚大气的包装简洁流畅的线条以紫色为主色调的搭配无一不显示出和天下的不凡紫色在中国传统里是尊贵的颜色如北京故宫又称为“紫禁城”亦有所谓“紫气东来”口感上此款雅香绵滑香气醇和余味纯净实属难得一见的神品啊.....	上海市	51.71421	积极	1
哪里能买到真烟？	湖北省	0	中性	0
淡，感觉除了淡还是淡，价格很离谱。国产烟的通病就是质量差价格虚高。。	浙江省	-3.21719	消极	-1
味道有点重第一次抽有点晕	安徽省	-2.48734	消极	-1
和天下这烟真的不行，味道还没软白沙好。	湖南省	-2.88194	消极	-1

## 4.2. 情感评分的可视化

首先看一下评论中消极、中性、积极情感的人数分布，如图 3。

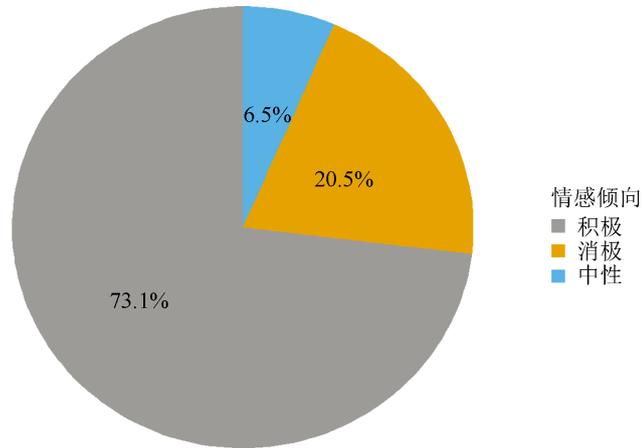


Figure 3. The proportion of people with different emotional tendencies

图 3. 不同情感倾向的人数占比

从上图可以看出，客户对白沙大都给予了好评，且好评的比例已经达到了 73.1%。接下来为消极评论占比 20.5%，最后为中性评论占比 6.5%。说明本产品是客户心目中的地位还是比较不错的。

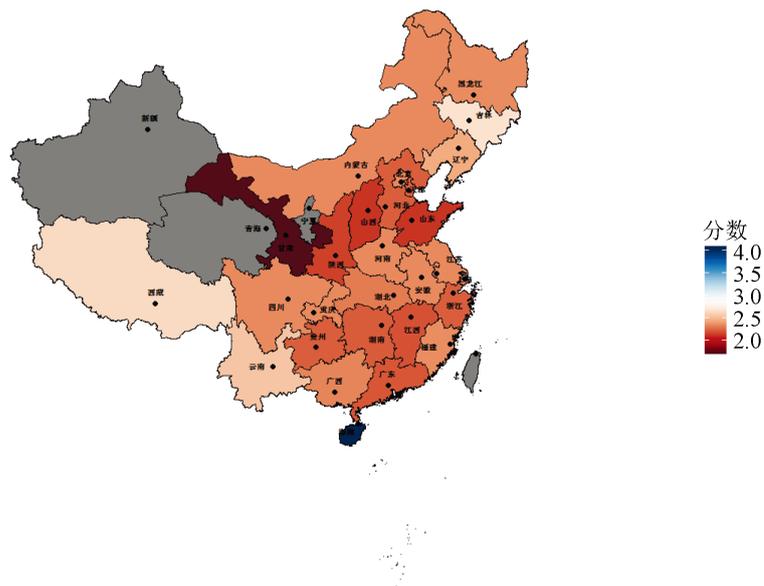


Figure 4. National emotional score thermal map

图 4. 全国情感评分热力图

从上图 4 可以看出，图例中颜色由红变白再变蓝的过程中得分逐渐增加。可以看出长江沿岸的省份对白沙的评价普遍稍高于其它省份，全国评分最低的省份是甘肃，最高的省份是海南。

## 5. 总结与展望

### 5.1. 总结

本文将现今较为成熟的文本分析应用到烟草文本评论中，通过对文本进行分词、情感极性分析等方法量化消费者对烟草产品的反馈信息，从而实现对产品的评分，该评分结果不仅能够帮助厂商决策人员

调整和改进产品，而且对消费者挑选烟草产品有着重要的参考价值。

## 5.2. 展望

本文的不足之处也较为明显，在进行情感评分的时候，进行了很多假设导致结果的准确率下降，首先假设了权值是线性叠加的，这在多数情况下都会成立，没有讨论非线性的情况。并且基于词典的情感分析效率不如基于机器学习的情感分析效率。

## 参考文献

- [1] 董日壮, 郭曙超. 网络爬虫的设计与实现[J]. 电脑知识与技术, 2014, 10(17): 3986-3988.
- [2] 中国互联网信息中心. <http://www.cnnic.net.cn/>
- [3] 王贵烽. 汽车文本评论的情感极性分析[D]: [硕士学位论文]. 北京: 首都经济贸易大学, 2018.
- [4] Hatzivassiloglou, V. and McKeown, K.R. (1997) Predicting the Semantic Orientation of Adjectives. *Proceedings of the EACL-1997, Madrid, 7-12 July 1997*, 174-181.
- [5] Turney, P. (2002) Thumbs up or Thumbs down: Semantic Orientation Applied to Unsupervised Classification of Reviews. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, Philadelphia, 7-12 July 2002, 417-424.
- [6] 朱嫣岚, 闵锦, 周雅倩, 等. 基于 HotNet 的词汇语义倾向计算[J]. 中文信息学报, 2006, 20(1): 14-20.
- [7] 徐琳宏, 林鸿飞, 杨志豪. 基于语义理解的文本倾向性识别机制[J]. 中文信息学报, 2007, 21(1): 96-100.
- [8] 闻彬, 何婷婷, 罗乐, 等. 基于语义理解的文本情感分类方法研究[J]. 计算机科学, 2010, 37(6): 261-264.
- [9] Pang, B. and Lee, L. (2005) Seeing Stars: Exploiting Class Relationships for Sentiment Categorization with Respect to Rating Scale. *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, Morristown, 25-30 June 2005, 115-124. <https://doi.org/10.3115/1219840.1219855>
- [10] 唐慧丰, 谭松波, 程学旗. 基于监督学习的中文情感分类技术比较研究[J]. 中文信息学报, 2007, 21(6): 88-94.
- [11] 伍星, 河中市, 黄永文. 基于弱监督学习的产品特征抽取[J]. 计算机工程, 2009, 35(13): 199-201.
- [12] 王献伟. 文本情感分析在商品评论中的应用研究[D]: [硕士学位论文]. 杭州: 浙江工商大学, 2018.
- [13] 周德翰, 李舟军. 高性能网络爬虫: 研究综述[J]. 计算机科学, 2009, 36(8): 26-29.
- [14] 周茜. 基于网络爬虫的信息采集分类系统设计与实现[D]: [硕士学位论文]. 厦门: 厦门大学, 2013.
- [15] 涂铭, 刘祥, 刘树春. Python 自然语言处理实战(核心技术与算法) [M]. 北京: 机械工业出版社, 2018: 38-58.
- [16] 儒小逸. 为什么 Python 适合写爬虫? [EB/OL]. <https://www.cnblogs.com/benzone/p/5854084.html>, 2016-09-08.
- [17] 於伟. 中文微博情感词典的构建研究与应用[D]: [硕士学位论文]. 上海: 上海师范大学, 2017.

### 知网检索的两种方式:

1. 打开知网页面 <http://kns.cnki.net/kns/brief/result.aspx?dbPrefix=WWJD>  
下拉列表框选择: [ISSN], 输入期刊 ISSN: 2169-2556, 即可查询
2. 打开知网首页 <http://cnki.net/>  
左侧“国际文献总库”进入, 输入文章标题, 即可查询

投稿请点击: <http://www.hanspub.org/Submission.aspx>

期刊邮箱: [ass@hanspub.org](mailto:ass@hanspub.org)