硕士学位论文

面向汽车在线评论的情感分类研究与应用

Research and Application about the Sentiment Classification of Automobiles' Online Reviews

黄 鹤

哈尔滨工业大学

2013 年 6 月

国内图书分类号：C931.6　　学校代码：10213
国际图书分类号：681.37　　密级：公开

硕士学位论文

面向汽车在线评论的情感分类研究与应用

硕士研究生　黄鹤
导　　师　叶强教授
申请学位　硕士
学　　科　管理科学与工程
所在单位　管理学院
答辩日期　2013 年 6 月
授予学位单位　哈尔滨工业大学

Classified Index: C931.6
U.D.C: 681.37

Dissertation for the Master's Degree

RESEARCH AND APPLICATION ABOUT THE SENTIMENT CLASSIFICATION OF AUTOMOBILES' ONLINE REVIEWS

Candidate: Huang He
Supervisor: Prof. Ye Qiang
Academic Degree Applied for: Master of Management
Speciality: Management science and engineering
Affiliation: School of Management
Date of Defence: June, 2013
Degree-Conferring-Institution: Harbin Institute of Technology

摘 要

纵观 Web 2.0 世界，用户原创内容(UGC, user-generated content)吸引了众多数据挖掘领域学者的目光，获取、跟踪并最大化利用这些用户原创内容也逐渐变成企业相关部门的一项任务。随着互联网技术的普及，与过去的十几年相比，对投放的产品和服务获取市场反馈越来越容易，越来越多的企业通过在线评论，获取市场反馈情报。因此，在线评论不仅是消费者购买决策的重要依据，同时也为企业所用，辅助与支持决策。进而如何提高 UGC 的投资回报率，使得工作更有效率，是每个企业关注的。汽车作为一种高价、不常购买的商品，比较而言消费者会倾向于把更真实的感观发布到网上，因而高效的识别出消费者对汽车的评论情感倾向并提炼出关键性的问题，对汽车企业有很大的应用意义。

本文旨在探究情感分类模型在汽车领域在线评论的效果，找出适合汽车在线评论的情感分类方法，编写汽车评论情感挖掘系统。首先，论文对国内外文本情感分类模型的研究现状做了系统性地总结，并归纳出三个主要的文本情感分类算法，朴素贝叶斯(Naïve Bayes)、支持向量机(SVM)、决策树 C4.5(J48)；本文为研究这三种算法对汽车评论情感分类的优劣，在论文实验部分抓取两个来源的汽车在线评论数据，将大量数据规范化预处理，导入情感分类模型，对三个算法的分类性能进行比较，最后利用调试出的最优模型对测试样本进行分类，编写汽车评论情感挖掘系统。

本研究可以高效地从海量评论中获得民众对于某一产品或服务的某些特征的正负面评价。在技术导向上，本文在基于机器语言的中文文本情感分类研究上做出定量分析，为汽车评论的情感识别建立汽车领域情感词典；行为导向上，本研究对汽车行业在线评论情感分析进行了深入的探讨，有很好的实践应用价值，为后续企业构建市场数据挖掘系统提供了一个开发的方向。

关键词：情感分类；用户原创内容；机器学习；汽车评论

Abstract

Across the Web 2.0 world, we’re seeing a quiet-but-knee-jerk shift away from UGC in favor of professional content, tracking and collecting data, making maximum use of this content is an assignment of relative enterprises. With the rapid development of Internet and informationization construction, the resources in the net increases exponentially, it's easilier to get the market feedback on Internent. That is, online reviews are not only utilized for consumer decision making, but also for the enterprise. However, how to increase return on investment of UGC, how to make more productive is the concern of every enterprise. Automobile, as a high-price, infrequently bought commodity, consumers would publish the real emotion online relatively, so the key issue here is to identify consumers' emotion inclination, it has the particularly important practical application value in the automotive field.

This paper is aiming at the classification result of each sentimental classifier on the online reviews of automobile field, finding out the best fit classification, and finally building and supporting a sentiment mining of Auto reviews system. At first, this article analyzes the reseach background and current situation of the text sentiment classification, and sums up three mainly models, Naive Bayes, SVM, decision tree to classify our data. In order to balance pros and cons of these three methods, in our experiment, our paper divides the automobiles' online reviews into two part, structured reviews and unstructured reviews. So we fetch the data from two different sources, aggregate and transform the data, and then import models, check and compare the relative index.

So this research application can efficiently obtain the positve and negative evaluation of the traits on the product or service, this can assist decision effectively. In the technical guidance, this paper gives quantitative analysis on the text categorization model based on machine learning, sets up a full-scale automobile sentiment dictionary; in the behavior orientation, this research discusses deeply in the sentimentic classification of online reviews in the auto industry, it possesses very important value in theory and practice, it gives a guidance on the building-up of integrated online market data mining system.

Keyword: sentiment classification, UGC, machine learning, automobile

目　录

第 1 章 绪　论

1.1 课题研究背景

近年来，随着互联网技术的进一步发展与普及，社会媒体改变了人们的生活与交流习惯，成为人们生活的重要组成部分。根据CNNIC中国互联网信息中心于 2013 年 1 月发布第 30 次中国互联网络发展状况报告统计，直至 2012 年底，国内总体网民数量达到 5.64 亿，互联网普及率为 42.1%[1]，保持低速增长。随着Internet技术与应用在过去十几年中的迅速发展，普通民众的生活习惯也随之发生巨大的改变。人与人的信息传递不再受地理位置和时间差异的限制，人们可以在任何地方、任何时间进行方便快捷的信息交流和情感沟通。此外，人们几乎可以在互联网上找到任何所需要的信息，人们还可以发布自己所拥有的信息供他人阅读。随着当今时代无线网络、移动互联网的迅猛发展，信息传播度和透明度也大大增加，用户浏览网站摄取信息的同时，也有分享发表个人看法的需求，因此，越来越多的用户愿意将真实的感观体会发布在网络上，并希望得到他人的反馈，从而商品的在线口碑迅速地发展并壮大了起来。而这些非结构化的信息产物就是UGC。

UGC 是用户原创内容(User-generated content)，亦称作 CGM(Consumer generated media)，泛指互联网上以任何形式由用户创作的信息，其表现形式可以包含文字、图片、音频视频等多媒体内容；这些内容与传统互联网信息的主要区别是，其最大的特征是由普通的互联网用户创造发布，他们不是来自于某组织企业的专职互联网信息发布者，这些发布的信息往往来自于个人真实的心得体会，或与互联网中其他普通用户的一种沟通互动。用户原创内容的典型代表网络平台有社交网络、论坛博客微博等，比较知名的国外网站有 Facebook 、Twitter 等，国内目前使用量和影响度较大的应用有人人网、优酷、微博、各大论坛等等。由于 UGC 的信息均来自于普通用户，相比于商家的广告宣传、推广服务信息等更具有用户主观体验真实性，所以用户原创内容的可信度和影响力是十分巨大的，对企业数据挖掘的价值也是巨大的。所以本研究主要针对于普通用户评论进行数据挖掘工作。

根据CNNIC发布的《2012 年中国网络购物市场研究报告》[2]可知：截至 2012 年 12 月 31 日，有 28.4%的用户表示自己最近半年使用过社会化分享网站分享购物心得，这些用户中最多的社会分享网站是微博类网站，有 62.6%的使用比

例。使用社会化分享网站的用户中，有 52.8%的人表示自己在社会化分享网站上浏览、关注过商品购物方面信息。在这些关注社会化分享网站上商品信息的用户中，有 18%的人在决定购买之前会经常登录社会化分享网站看看相关商品信息，分别有 25%和 37%的用户有时或偶尔登录查看。说明社会化分享网站的相关信息将得到越来越多的欲购者的关注，用户越关注，企业就越需要及时掌握，因此本研究以收集微博用户评论数据为主，收集新浪汽车网站的评论数据为辅做以比较分析。

随着我国经济的飞速发展和人民生活水平的显著提高，汽车已经成为人们工作和生活中的必需品。据中国汽车工业协会统计分析，2012 年全国汽车产销 1927.18 万辆和 1930.64 万辆，产销突破 1900 万辆创历史新高，再次刷新全球记录，连续四年蝉联世界第一[3]。随着汽车的不断普及，拥有一辆汽车已经成为现代家庭生活的一项标志。因此，越来越多的消费者计划购买汽车。但由于经济条件的区别或者价值观的差异，中国的消费者购车价位从几万到几十万，甚至到上百万。根据搜狐汽车网统计的数据，2012 年我国汽车销量为 1930.64 万辆，其中轿车(主要包括小型车、紧凑型和中型车)销量最多，占 52.6%；货车占 15.7%；交叉型乘用车(微客，俗称面包车)占 13.8%；SUV(Sport Utility Vehicle，俗称四轮驱动越野车)占 7.3%[4]。根据易车网 2012 年的汽车统计数据可知，在轿车、交叉型乘用车和SUV车型中，消费者购买主要集中在中、下档汽车[5]。以上数据表明，汽车购买力主要来自在普通群众，购买的主要是中低档价位的汽车，且购买意图主要是家用、办公等。

汽车的口碑对有购买需要的消费者来说显得尤为重要，因此消费者策划购买汽车前查找相关资料、了解汽车性能尤其是查看汽车评论是必不可少的。网络信息不但快捷、即时、覆盖范围广、准确，而且网络信息具有多元化的特点，消费者不仅可以看到相关产品的介绍，还可以看到其他消费者对该产品的使用后的真实体会经验。中国领先的网络口碑(IWOM)研究咨询公司 CIC 公司研究中发现，汽车口碑网站的用户活跃度和互动性非常高，每月有超过一千万条评论产生，其中包括很多针对各类车型的用户反馈，从购车咨询到试驾、购买体验，从购买后体验到经验分享，一应俱全；同时，各类针对 4S 店服务的投诉建议类信息也较多。因此，企业有需求通过数据挖掘系统分析挖掘出包括竞争车型在内的各类产品体验反馈，并通过指标量化各个反馈内容，为企业提供关于产品和服务反馈的定量定性数据，帮助企业决策人员调整相应的服务改进体系。

目前，我国已经有很多专业的汽车网站提供丰富的汽车信息以及汽车在线口

碑信息，为用户提供发表在线评论的广阔的平台，比如太平洋汽车网、汽车点评网、网上车市等。据中国排名数据网站(http://top.chinaz.com/)最新数据统计显示，生活服务类网站按照得分排名，据统计在生活服务类网站中，包括求职类、房产类、汽车类、金融银行类、家居美食类等网站，60 名内，汽车网站有 12 家，占总比例 20%，说明汽车类网站的发展速度是非常快的，用户也越来越需要汽车信息覆盖全面、全方位测评的优质汽车类网站，而这类汽车网站在受到用户青睐的同时，企业及时在网站中发掘价值信息，了解市场趋势，对企业汽车研发及创新是非常有必要的。其中，新浪汽车网作为ALEXA排名 1第 17 位居汽车网站榜首。一些重要的门户网站也开辟了类似的汽车专栏，内容丰富度堪比专业汽车网站，比如新浪汽车网、搜狐汽车网和网易汽车网。由于门户网站的知名度较高，其汽车网也得到了网友们的广泛关注，他们积极的参与汽车网的在线评论，使门户网站的汽车网迅速发展壮大起来。此外，目前还存在着大量的汽车网上论坛和在线车友俱乐部等供车友们进行交流。在中国零星已经出现了汽车网购的苗头，还未得到普遍大众的接受和参与，目前大部分消费者只会在网上购买汽车的相关配套设备，如坐垫、地毯等。但是随着科技的进步和社会的发展，人们的观念也在不断地改变，高价值产品网上销售也将成为可能甚至是必然。因而网上销售汽车指日可待，这预示着在不久的未来汽车在线评论将会变得越来越重要。

而对于社交类网站中，新浪微博作为近几年新兴的社交类网站，以 ALEXA 排名 28、百度权重 9 分，谷歌 PR8 分，综合得分 4380 居于社交类网站第一名。2012 年用户访问微博的频率和时间百分比远超过社交媒体网站，因此，目前微博无疑是一个不可忽略的信息发布和收集平台，微博近几年的快速发展，潜移默化地影响普通民众的生活习惯。微博作为一个大环境下的传播媒介，涵盖各个领域的松散信息，汽车类微博也不例外，也属于汽车评论的范畴，但是汽车类微博一定与传统汽车评论存在某种联系，也存在一定的区别，作为一个创新性社交媒介，对这类评论对管理科学研究人员来说，更具有探索价值。本文将对其进行区分并比较研究。

因此，如何利用高效利用汽车在线评论这一潜大市场信息模块增加企业的竞争力和业绩,将会是汽车行业线上监测影响线下销售的一个创新的探索过程。

[1]Alexa 排名：Alexa 中国免费提供 Alexa 中文排名官方数据查询,网站访问量查询,网站浏览量查询,排名变化趋势数据查询。

1.2 国内外研究现状及分析

对于在线评论的情感分析研究工作近几年发展迅速，国内外学者已经有了很显著的研究成果。

在国外的研究中，学者对于文本情感倾向性挖掘的研究工作开展比较早，涉及领域也相对比较多，相关文献中可以找到关于移动电话、电影票房、旅游酒店等产品或者服务的在线评论情感分析研究。其分析方法主要以单词为单位粒度展开，简单的说，是从数据样本中提取出的情感关键词或者对应领域的相关的固定词组，将这些提取出的词和词组组成一个情感词列表，分为正向情感词列表和负向情感词列表，并查找文本中存在的列表中词汇，并计算相应的特征值，最后参考所有值识别出该文本的情感倾向。

在金融领域中，Mike Y. Chen[6]等人对雅虎网站股票的留言板信息数据进行提取，研究股票投资者的态度。Bo Pang 等学者在 2002 年的研究中应用机器语言学习方法对文本式评论进行情感分析[7]。日本电器股份公司也在对服务关注度信息进行语义提取并在情感倾向性的工作中取得了一定研究成果[9]。Philip Beineke 等研究者在用机器语言学习过程的同时，还加入人共注释评论一些工作，以提升了英文文本评论情感分类识别的精准率[8]。Turney等人在 2002 年发表论文中提出了基于语义空间模型(SO，Semantic Orientation)的情感分类研究方法，利用PMI-IR算法对数据进行情感分类，对汽车、银行、电影和旅游目的地等商品和服务的客户评论进行研究，得到汽车评论的准确率是 84%，电影评论的准确率是 66%，显示出这类方法已具有了初步的应用价值[10]。

国内方面，孙文俊等采用情感分析与统计学方法结合，以网上图书评论数据为研究对象，从评论的文本特征出发，构建在线评论有用性影响因素模型[11]。同时，S Wenjun, P Mingyang, Y Qiang还研究了来自不同来源的在线评论，主观和客观评论的所含情感不同且目标读者也有所区别，并为卖家提出战略性建议[12]。丁宇新[13]等学者采用朴素贝叶斯模型和最大熵模型对新闻评论进行分类的研究，分类过程中尝试两种计算方式，计算情感词频率和二值作为特征权重，实验结果表明采用二值为特征权重的分类效果比较好，利用同种特征选择方法时，最大熵模型的情感分类效果最好。林鸿飞等学者将构建语料词典和机器语言学习方法相结合，首先对文本提取褒贬情感倾向词、程度性副词和否定词，对这三种词汇依次计算特征值，接下来采用支持向量机分类器和加权计算词频两种方法，比较对样本数据进行褒贬分类的结果[14]。

而对于微博的情感分析国内外的研究发现，国外研究方面，Go, Pak A和

Paroubek P等学者较早对Twitter这个国外著名的微博网站开展情感分析研究工作[15]。Go等人采用 NB(朴素贝叶斯)、SVM (支持向量机)[16]、CRF(随机条件域)[17]对tweets短文本情感分类，实验结果得到支持向量机的结果最佳。在特征提取过程中，采用构建词典、二元分类模型组合POS(parts-of-speech)进行特征提取，结论得到词典构建这种特征提取方法对于tweets的处理最终分类器得到的结果最好。在这基础上，Pak等学者对tweets分类工作有所突破，将情感分类过程分成两步实现：主观性文本识别和情感极性文本分类。Feng J等根据tweets短文本的特殊语法特点，将转发、话题标签(##)、短链接、标点符号、英文表情、感叹标志等因素考虑在内，结果表明这些特有语法特征对情感分类产生的影响不容忽视[18]。综上所述，国外学者针对twitter英文博文的研究成果目前已经比较显著。而在国内研究中，现有中文情感分析研究关注点都主要集中在电影评论、产品在线评论、旅游酒店评论等，由于微博这类信息发布平台是最近几年新兴的热点社会媒体发布平台，因此国内该领域内研究学者对于微博消息情感倾向性分析研究相对还比较少。

1.3 研究内容与研究意义

本文主要的研究重点在于对汽车在线评论进行情感分类研究，首先，以新浪汽车网在线评论口碑作为传统汽车评论研究案例，对其进行情感分类研究，然后进行微博的情感分类，以新浪微博作为第三方数据来源。将二者研究结果和研究过程进行对比参照，识别出针对汽车在线评论情感分类的最优模型。本系统在 Windows 7 操作系统，采用 Java 语言开发(JDK 版本为 1.6)。整个系统体系流程包括数据采集整理、语料规范化预处理、中文分词、特征值提取、分类模型的选用、可视化结果的展示。

论文从以下几个方面进行研究：

(1) 通过研究结构化汽车评论的特点，改进本研究的情感词典，先找到适用于网络语言挖掘的情感词典，并找到适应于汽车领域自身特点的情感特征词集合，构建针对汽车行业的分词词库，不断改善导入分类模型文本质量；

(2) 在传统利用机器学习进行文本分类的基础上，本研究针对结构化汽车评论和非结构化汽车评论均采用三种普遍运用的情感分类模型进行分类，在证明机器语言分类器有效性的同时，可以比较哪种机器语言更适合对本实验样本进行分类，是否结构化评论和非结构化评论的分类效果存在差异；

(3) 开发针对汽车评论的情感挖掘系统，通过对采集样本进行自动分类并提取，可视化将结果以人机交互界面的形式呈现，辅助企业市场人员和决策人

员高效进行市场分析调研工作。

本文的研究意义主要有：

(1) 本文将挖掘汽车资讯网站及微博中的汽车在线评论，并且概括顾客们对各种汽车车型的不同性能指标或重要部件的评论和意见。

(2) 通过不同来源的样本数据探究对于分类模型的性能优劣，采用三种机器文本分类方法对评论文本进行分类，并调试出最优的情感分类模型，自动判断评论的褒贬性。

(3) 构建汽车评论情感挖掘系统，实现利用自然语言处理技术自动化将汽车评论本体分类，并抽取出每类情感的关键标签，从而帮助企业快速直观从海量市场反馈中提炼出有价值的信息，更准确更快速的做出全局统计，从而支持战略决策。

对于汽车在线评论的情感分类，只是在线评论情感分析的一个领域。所以，本研究将在探索在线评论情感分析方法的基础上加以改进，建立适用于汽车行业的评论情感分类模型，由于汽车目前还属于线下销售商品，与普通的网购商品存在差异，且汽车与人们的生活息息相关，所以本文将以汽车评论为研究对象，探讨汽车在线评论的情感分类模型，构建汽车评论情感识别系统，把握用户的情感走向，探究高效准确的汽车在线评论情感识别方法，从而帮助汽车企业了解消费者心理，改善营销策略，并最终提升企业的形象、提高企业的利润。

在技术导向上，本文在基于机器语言的中文微博情感分类研究上做出定量分析，为汽车评论的情感识别建立汽车领域情感词典；在行为导向，本研究对汽车行业线上评论情感分析进行了深入的探讨，有很好的实践应用价值，为后续社会媒体数据挖掘系统地建立提供了一个开发的方向。

1.4 论文组织结构

基于上述研究内容，本论文共分为五章展开，每章的研究内容如下：

第一章 绪论。介绍课题的研究背景、研究目的和研究意义。并对国内外研究现状进行综述，最后提出本文的主要研究内容和论文结构。

第二章 介绍本文所涉及的相关基础理论。包括在线评论情感分析基础理论和微博情感分析的理论。通过阐述在线评论和情感分析的相关理论知识，为之后的实验研究做好准备。

第三章 结构化汽车在线评论的情感分类实验。编写 Java 程序来收集结构化汽车网络数据来源的样本，阐述收集数据的方法、数据整理及语料预处理过程，对预处理后的文本采用三种机器语言模型进行情感分类，得到三种机器语

言模型的召回率和精确率，对三种机器语言模型分类结果进行对比分析，

第四章 非结构化汽车在线评论的情感分类实验。定义数据抓取规则及边界，编写微博网络爬虫程序采集非结构化并对数据进行编译，同样对语料规范化预处理、描述性统计分析，导入分类器进行结果讨论。

第五章 探讨两种评论的情感分类过程的异同,利用最优算法构建汽车评论情感挖掘系统，建立情感抽取模块，可视化显示汽车在线评论褒贬分类的抽取出的关键词，以文字云的形式展现。

结论，总结本文所取得的研究成果，并指出课题研究中目前存在的不足和对未来研究工作开展的思路，并对汽车行业的在线评论研究前景作了分析。

本论文的研究路线框架图如图 1-1 所示，研究将汽车在线评论分为结构化汽车在线评论与非结构化汽车在线评论两种评论进行对比研究。其中，框架中左半边虚线框中是本文第三章的内容，右边的虚线框中是研究第四章的内容，

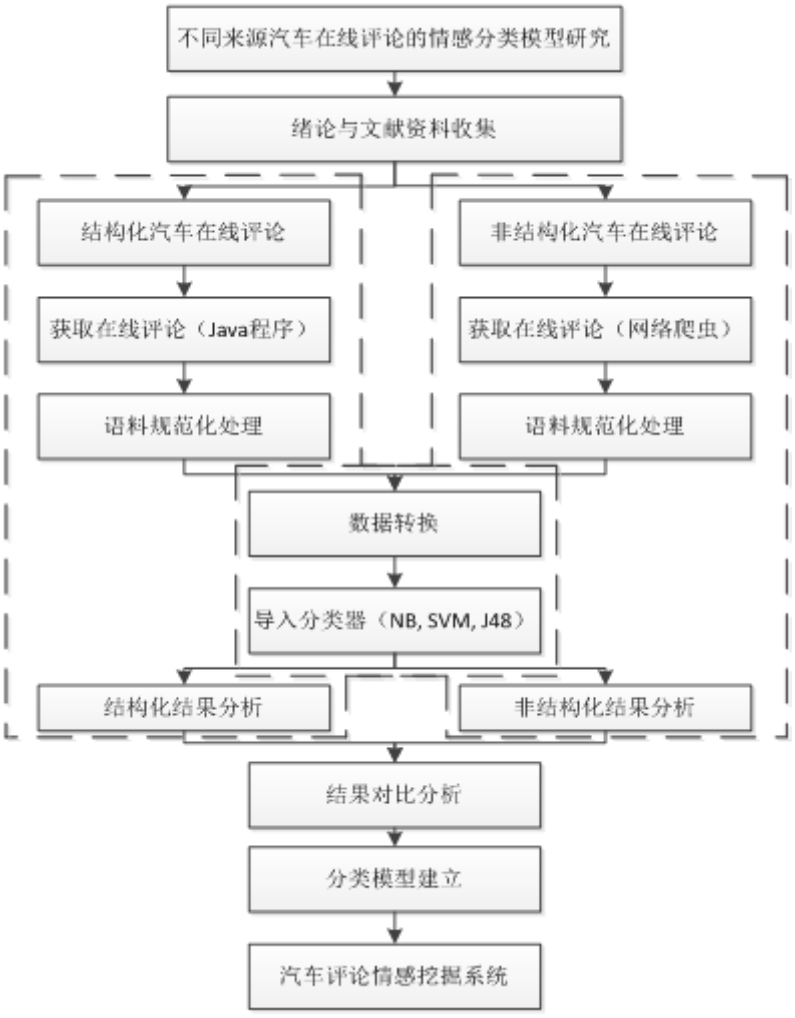


图 1-1 研究路线框架图

第 2 章 基础理论与相关技术

本章主要阐述论文的理论基础，包括在线评论的理论基础和情感分析的理论基础。首先，本章介绍了在线评论的概念、特点，然后通过对比在线评论与在线口碑之间的关系来进一步说明在线评论的特点，作为本文研究的一般基础理论。然后，接下来的小节介绍了情感分类技术的相关理论。这些理论为本文在以下章节提出三种机器语言分类模型支持，并为本文汽车评论情感挖掘系统的模块构建提供依据。

2.1 在线评论理论基础

在线评论，是用户原创内容(UGC)很重要的一部分，指普通用户在网络上针对某事物的评论信息。随着网络形式的不断丰富，在线评论的表现形式也变得多种多样，其大致可分为两种形式，即结构化良好的在线评论和非结构化化在线评论。

Liu B, Hu M和Cheng J提出在线评论按结构可分为三类：①评论分别列出商品的优点和缺点；②评论列出优点和缺点，同时可以进行自由评论；③无固定格式的自由评论[19, 20]。可以看出，前两种都是结构化良好的在线评论，结构化评论的特点是大都比较规范化，除了基本的文字性内容的评论，还引导消费者对产品或服务进行评分或评级，甚至引导消费者对产品的各方面属性都进行量化的评价，其发布形式都应以格式化的文字或评分存在，这类评论在数据收集整理时相对容易，评分或者好评、中评、差评的类别即可作为文本性评论分类的参考，这个量化的指标在一定程度上可以成为评论情感倾向的代理，非常有利于进一步的量化分析。比如各大网络商城(如淘宝商场、亚马逊商场、京东商城等)、团购网站(如拉手网、窝窝团等)，还存在于一些专门的在线评论网站(如大众点评网)，此外，结构化在线评论还存在一些资讯网站上的某个板块中，比如本文要研究的新浪汽车网；非结构化评论有量化比较困难的特点，这些评论都是用户们针对某个事物发的感观体会，比如，网上论坛(如汽车之车，为汽车爱好者提供交流互动的平台)、虚拟社区、网络博客和微博(如新浪微博，用户可以表达对各领域各话题的体会和他人、企业、名人互动)等网络空间的评论，都属于非结构化评论，因此，这类评论不仅没有量化的打分机制作为参考指标，而且在评论数据抓取和筛选中也存在较大的困难。

综上所述，结合Bickart和Schindler对在线评论的各种表现形式及其特征的描述[23,24]，和郝媛媛[25]在文章中对在线评论的总结，本研究对在线评论的多种形式做了如下说明归纳(见表 2-1)。

表 2-1 在线评论的分类说明

类型	特点	举例	本文研究案例
结构化在线评论	常持续一年以上，有星级评分，结构化良好，列出优缺点的同时，有的还有自由点评	淘宝网等网购商城，团购网，点评网站，旅游网站，资讯门户网站	新浪汽车网
非结构化在线评论	亦可持续相当长时间，传播方便、快捷，结构自由松散，存在表情、图片、多媒体、分享链接等多元化信息	虚拟社交网站，论坛帖子及回复，即时通讯聊天记录，微博	新浪微博

总之，无论是哪种类型的在线评论，最主要的差异体现在形式和存在空间的区别，但是它们的核心内涵是不变的，即都是消费者针对某产品或者服务的非正式的个人心得体会的表达，由于普通用户发布的在线评论都是根据自己的经验，这些经验往往是表达体验后产品或服务的性质和质量。而专业机构等在第三方网站发布的评论，往往为市场推广营销作用，为赚取社会声誉，这些信息都是中立的，他们的评论更关注商品的属性本身而不是体验后的感受，因此这些专业机构评论不存在情感色彩。如制造商和分销商等这些专业机构发布的网上评论，一部分是对产品或服务的客观属性描述，另一部分是营销广告。因为对这一类广告的甄别意义不大，所以不是汽车情感挖掘系统关注的目标数据集。由于互联网的特性，这些网上发布传播的用户原创内容都有历史数据保存完整的特点，这方便于开展量化分析。

本文的主旨是为企业快速识别用户的体验感受，所以研究对象是普通用户发布的结构化评论和非结构化评论并展开研究。以下小节将基于在线评论的一般性情感分析技术来讨论在线评论的情感分析理论基础。

2.2 情感分类技术理论基础

2.2.1 情感分类概念及相关理论

情感分析兴起不久，是近几年文本挖掘和自然语言处理领域热门的讨论研究话题，受到了国内国外学者非常广泛的关注。情感分析是指通过计算机技术，抓取和处理文本、图像、视音频等对象所包含的情感内容，帮助用户或研究人员快捷简便的获得有利用价值的情感倾向及其强度等信息，如试图找出“喜欢”、

“不喜欢”的，或是一种“中立”的态度。情感分析研究按照研究粒度大小可以分为对篇章级别的分类，对句子级别，对词或短语级别等子问题；按照功能类型进行类别区分，分为情感分类，情感检索，情感抽取等子问题[21]。本文主要采取词或短语级别的粒度，通过情感分类，情感检索，情感抽取及情感数据的可信度问题来分析解决汽车在线评论的相关情感分类研究问题。本小节介绍了情感分类在情感分析领域的研究理论基础，情感分析研究体系结构如图 2-1 所示。

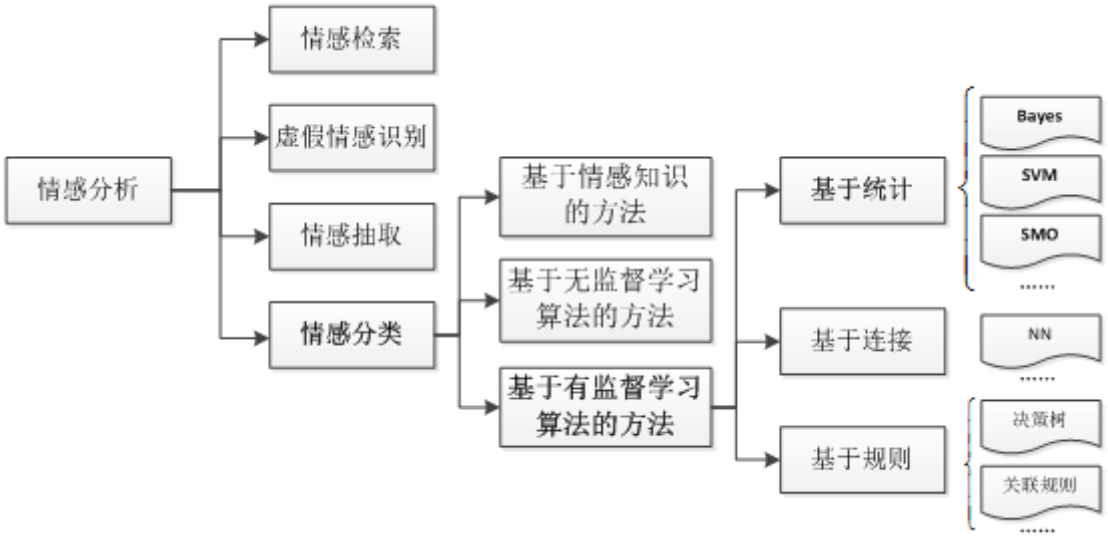


图 2-1 情感分析研究结构体系图

对于文本的情感分类研究，是对于一个给定的带有主观情感色彩的文本，识别出该文本的情感倾向，有时包括程度。一般对本文的情感分类研究是识别出给定样本的褒义文本和贬义文本。除了褒贬情感辨识之外，有的研究学者将情感分类分为三种类型：正面的、负面的和中性的，中性文本指没有用户主观明显情感倾向的文本。还有的研究则是将情感类别分为多种，对样本集中的评论进行打分，比如 1 到 5 分，则识别出了文本的情感程度。因此说情感分类不单单可以描述情感倾向，也可以描述情感的趋势程度，甚至是表达褒义情感的程度或表达贬义情感的程度。

目前的研究工作，大部分的工作都是讲文本进行褒贬区分，介于微博评论中存在大量客观宣传和描述，没有个人主观感情色彩，这类往往被归类于客观文本的范畴。因此，本文对于结构化评论采用褒贬中性三元情感分类，而对于微博，非结构化评论先进行主客观分类，再对主观情感分类进行褒贬性二元情感分类。

目前情感分类技术主要有三种技术导向研究：基于情感词典的文本情感分

类方法，无监督学习方法的文本情感分类和有监督学习方法。在目前的情感分类研究中，按照褒贬情感两个类别进行分类居多。无监督学习的本文情感分类方法是指使用不提前进行类别标注的文本集来识别文本集的情感类别，有监督学习算法则是指利用人工标注类别的训练数据集先进行训练，然后利用训练好的分类器，分析文本的情感类别。因为本研究主要结合了情感词典的分类技术和有监督分类方法中的一些机器学习方法，对无监督学习方法不做赘述。

（1）基于词典的情感分类方法

基于词典的情感分类方法，又被称为情感词加权方法，主要依据专家词典的基准词判定为基础准则。在文本处理的研究中，情感词表的获取，对于英文词表，大多研究普遍基于第一章提到的普林斯顿大学开发的 WordNet，它是一个大型英语词汇汇总数据库，在 WordNet 中，名词，动词，形容词和副词各自被统计成一个认知相同意思的集合，每个同义词汇集合都各自代表一个基本的语义概念，且各种关系链接也存在于各个集合之间。每个集合都有自身蕴涵关系的层次，如名词中的上下位关系，这样就组成了一个同义词的网。

另外， Peter Turney[10]在情感基准词的基础上，利用点互信息量(Point of Mutual Information，PMI)对文本进行情感倾向计算。在基于点互信息量的文本处理方式中，假设存在一些具有已知情感极性的基准词，通过词语之间的紧密关系程度来对词语的情感倾向性进行推断。也就是说，它使用点互信息量来衡量词语的情感倾向性，这需要两部来完成对情感倾向的处理：第一步，抽取情感短语，第二步，计算词语的情感倾向。

而对于中文文本处理上，依靠的词典来源往往是知网和同义词词林的扩展版。这种基于中文词典的主旨是：给定一组已知极性的基准词集合，对于文本中的情感倾向未知的词语，比较其与基准词的语义相似度，划分阈值，识别其情感倾向性，再经过文本中正负情感的次数，根据公式(2-1)得出文本的情感倾向：

$$polarity=\begin{cases} Positive\ (if \quad posCnt > negCnt) \\ Neutral\ (if \quad posCnt = negCnt) \\ Negtive\ (if \quad posCnt < negCnt) \end{cases} \tag{2-1}$$

利用情感词典分类，尽管简单、准确率高，但分类的好坏完全取决于情感词典的质量，如果文本语义结构复杂性较高时，单纯地利用情感词典的词频计算文本情感极性，并不能取得良好的分类效果，需要考虑文本的上下文情景、句法结构、修辞手段等因素对情感分类造成的影响。

（2）基于有监督学习算法的情感分类

有监督的机器学习模型基本上又可以分为三大类。一种是基于统计的方法，如贝叶斯、k 最近邻分类(k-Nearest Neighbor，简称 KNN)、支持向量机(SVM)等方法；另一种是基于连接的方法，即人工神经网络；还有一种是基于规则的方法，如决策树、关联规则等，这些方法的主要区别在于规则获取方法。而它们之间有个共同点，即都需要先对模型进行训练，形成训练器再对其进行测试。本研究就是基于机器学习算法的基础上，对汽车评论样本进行情感分类研究。

B. Pang在研究中利用了机器语言学习模型定义电影评论的情感极性，他们首先提出了文本预处理的概念，包含消极词汇提取、单词提取、多词提取、定义标签、提取倾向性信息的概念，并且作为特征表示，他们利用朴素贝叶斯(Naive Bayes)，KNN和SVM算法进行情感分析[31]。结果显示，在影评数据中，选取单词特征值时，SVM获得最好的分类效果，精确度达到 83%，其次是NB和ME。Kim不仅在分类中引入n-gram模型外[32]，同时考察了位置特征和评价词特征对句子级情感分类的影响。本研究将对汽车评论样本利用朴素贝叶斯，SVM，决策树C4.5。

1) 朴素贝叶斯算法

介于朴素贝叶斯算法(Naïve Bayes, NB)的简便性和快捷性，它经常运用于文本分类。朴素贝叶斯算法(NB)，是一个基于贝叶斯理论的概率学习方法，是根据事件发生过的频率分布进行分类预测的方法，可以表达随机变量之间的复杂概率分布。它的本质是利用类别的先验概率和类别的条件概率，通过贝叶斯公式得出未知文档的概率。

假设一个给定的文本 d 属于类别 c 的概率为，

$$P(c|d) \propto P(c) \prod_{1 \le k \le n_d} P(t_k|c) \qquad (2\text{-}2)$$

其中，$P(t_k|c)$—— 条件概率；

$P(c)$—— 文本 d 属于类别 c 的先验概率；

t_k—— d 对应的 k 维特征。

在文本分类中，目标是找到文本属于的最佳类别，利用NB分类找到的最佳类是计算最大后验概率(MAP, maximum a posteriori)c_{map}的过程，

$$c_{map} = \arg\max \hat{P}(c|d) = \arg\max \hat{P}(c) \prod_{1 \le k \le n_d} P(t_k|c) \qquad (2\text{-}3)$$

朴素贝叶斯必须符合前提假设所有属性都是相互独立的，数值属性需要是离散的。这往往限制了朴素贝叶斯的准确性。

朴素贝叶斯分类算法简单高效，在对待具有不同数据特点的数据集合时，分类性能的差别不大，所以针对某种特点的数据集稳定性较强。尽管在实际情

况中，属性间类条件独立的情况很少，但是在大多数情况下，该分类效果仍比较精确。其主要原因在于需要估计的参数比较少，虽然概率是有偏估计的，但研究者们看重的恰恰不是是绝对值而是排列次序。对于存在缺损的数据，大多数情况下都会在使用前进行预处理，以此提高数据的可行性。

朴素贝叶斯分类算法也存在一些缺陷。前面提到，朴素贝叶斯算法要求条件属性间满足相互独立假设，这在实际情况中经常被违背。如果在实际的数据间存在属性间高度相关的情况时，会导致分类效果不理想。

另外，朴素贝叶斯分类算法从训练集中得到类别先验概率和特征项的类条件概率来计算预测后验概率，再根据公式得出需要被分类的实例的类别情况。如果训练数据样本量有限，则很难完全代表总体情况的分布，因此计算得到的先验概率的可信度存在问题。如果训练集没有良好的数据完备性，那么预测的测试数据的类别标签就可能不准确。

2) SVM 算法

与朴素贝叶斯相似，SVM方法也普遍应用于传统文本分类研究中，且在主观情感识别领域也发挥了十分广泛的作用[31, 33]，SVM 算法作为新兴机器学习算法，近年来被成功地应用到很多模式识别问题中，其在数学上表示为求解一个二次规划问题[34]。在本文的实验设计过程中，在测试SVM算法过程时设置了和B. Pang等学者研究中同样的参数设置。

SVM 算法，每个文本即是一个向量，每个坐标即是一个单词，通常情况下，单位长度向量需要归一化。SVM 算法的主旨是尽力在一个多维空间中寻找到一个满足分类要求的最优多维线性平面对文本进行分割，既可以保证分类精度，又使得分割平面的两侧的空白区域的宽度值尽量最大化。SVM 算法需要满足两个前提，前提一，文本在同一个类中组成一个连续的空间区域，前提二，属于不同类中的文本不重叠。它的应用过程中还引入了线性不可分问题处理的核函数，将低维空间的非线性数据通过非线性映射函数映射到高维空间中去。因此 SVM 算法不但解决了非线性分类问题，又避免了产生高维计算困难的问题。

SVM 也适用于多元分类应用领域，以其中的一类为正例，其它类作为反例进行建模，得到该类的最优分类面，剩下的其它类别都可以按照上述方法来进行建模，这样得到多个类别的二元分类器，最后把它们共同组成一个多元分类器。在下一章，我们将对 SVM 模型的应用进行深入的探讨。

3) C4.5 决策树

决策树方法最早产生于上世纪 60 年代，到 70 年代末。最早的决策树算法

为CLS，于 1966 年由Joyce B R等人提出[37]。目前研究中，最具有影响力的决策树算法是Quinlan J R分别于 1986 年提出的ID3[38]和 1993 年提出的C4.5[39]。其中，ID3 算法构造树的基本想法是随着树深度的增加，节点的熵迅速地降低，从而提高算法的运算速度和精确度，其特点是只能解决离散型描述的数据。

C4.5 决策树在 ID3 决策树的基础之上稍作改进，C4.5 克服了 ID3 的 2 个缺点：

1. 用信息增益选择属性时偏向于选择分枝比较多的属性值，即取值多的属性。

2. 不能处理连贯属性。

C4.5 采用了 IG 思想作为选择分枝参数的标规则，弥补了 ID3 算法的不足。

在归纳学习中，它代表着基于决策树(Decision Tree)的方法，通过对一组训练数据的学习，构造出决策树形式的知识表示，在决策树的内部结点进行属性值的比较并根据不同的属性值判断从该结点向下的分支，在决策树叶结点得到结论。

决策树算法的优点如下：(1)生成的训练模型可视化简单，易于理解；(2)分类精度高；(3)对于噪声数据、数据表示不当有很好的健壮性。因为这些优点存在，是目前应用最为广泛的归纳推理数据处理算法之一，在数据挖掘领域中受到研究者的关注。也就是说，决策树算法的最大优点是在训练过程中不需要用户掌握足够的决策树背景。因此，只要训练事例能够用属性的结构表达出来，就可以利用该算法来进行工作[40]。

决策树 C4.5 的缺点：在构建训练模型决策树的过程中，需要对样本集合进行反复的顺序扫描和迭代排序工作，从而降低算法的效率。与此同时，C4.5 决策树只适用于特殊的数据集合，也就是可以驻留在内存中的数据集，当训练样本集合很大时，会导致内存无法容纳，从而程序无法正常运转。

2.2.2 情感分类过程中关键技术

(1) 中文分词技术

在汉语文本中，词是能够单独表示意义的最小单元。利用计算机技术，对汉语分类并处理，是中文信息处理所需要解决的工作。由于汉语中不存在很明显的分隔标示来将不同的词分隔开，所以，需要将汉语字串进行切分处理，将字串单元处理成词串单元。想要对文本自动分析，首先需要进行自动分词，在这个过程中，可能面临以下的情况：在英文中，词在识别的过程中，把所有空格符号删掉后，处理器再对空格符进行自动还原，会出现很多歧义，需要对其

进行相关处理。中文和英文的差别此时就被切词所体现出来。体现的现象为，英文基于小字符集，词串完全分割。中文基于大字符集，字串为连续的形式。也就是说，导入到分词系统的的输入集是连续的字符串，而输出级则是一系列词串，这个词串既可以是一个单词，也可以是固定搭配的词组。

这种机械式的分词方式是基于词库的，遵循匹配的策略，把待分析的字符串与字库中的信息进行一种匹配。匹配可以分为不同的方式，基于匹配方向，分为正向和反向两种匹配方式；基于长度优选原则，分为最长和最短匹配方式；基于词性标注的结合度，分为单纯分词方法和一体化方法两种方式。

通常情况下，我们不采用正向反向匹配的方式。数据分析的结果显示，反向匹配的精度效果相对好一些，但是单单使用一种方法，仍无法达到实际需要的标准。所以，我们在进行分词工作时，一般只把这种机械性的手段做为初始的手段，若要提高准确程度，还必须结合其他手段，进行后续精确分词。

改进扫描的方法也是在分词中常用的一种手段。该方法的原理是在需要分析的字串中，识别具备显著特点的词，对其进行切分，并将切分出来的词列为断点词汇，把原来较长的字符串切割成小的单元，再展开机械式的分词手段，从而降低匹配出错的概率。另外，把分词以及标注进行结合操作，也可以在一定程度上增大切分的正确概率。

本研究使用的是中科院计算所开发的汉语词法分析系统(ICTCLAS3)，就是基于上述主旨开发。中科院计算研究所的研究表明，利用该所研发的这套系统，可以将分词的准确概率上升到 97.58%，这个准确率是由 973 名专家测评得出，未登陆词识别召回率也十分明显，分词的处理速度也有十分显著的提升。

(2) 排除停用词

有一类词语，在语言中会经常出现，这类词语实际上是可以被忽略的，我们称这类词语为停用词。因为它们虽然是构成句子的成分词语，却无法表达文本的意思。冠词、介词、连词、代词都可以被划分为停用词，比如“了”、“我们”、“至于”、等。

词语在做分类的时候，有如下两种词语，一种是功能词，这类词与文本要表达的内容并不相干；一种是对分类没有太大用处的词，原因可能是因为这类词在文本中出现频率太高或者太低。以上两种词语，我们称之为停用词。

在实际过程中，基于词频去除停用词的方法相对来说比较难实现，因此，我们用更加有效的方法——停用词表法来去除停用词。把具有高 DF 值和低 IDF 值、对文本贡献率很低的高频停用词整理出来，形成停用词表。理论研究表明，在英语中，高频功能词所占的比例在百分之四十与百分之五十之间。

除了以上两种词语，有些常用的词语也被编入停用词表，数量大约为 280 个。对于停用词的选择方法，中文与英文有很大差别，相对来说，英文比中文简单一些，因为英文很多时候并不像中文那样，在语言结构中有很多种含义，扮演的角色也可能多种多样。虽然如此，两种语言选择停用词的基本原理是相同的。很多时候，基于停用词表可以剔除掉文本中近一半的词汇，令后续分类相关工作更便捷。基于词的类别及含义，可以选出 1000 多的词语输入停用词表。

(3) 特征提取技术

巨大数量的词汇构成了文本，导致文本向量空间维数过于复杂繁琐。因此，维数压缩的工作是很有必要的，其目的是为了使程序运行效率增加并使分类精度得到明显提高。词语对文本分类的作用和贡献程度并不相同，有些对分类的作用很小，比如某些各个类别都广泛存在的词语、而经常在特定环境下出现，但在一般情况下很少出现的词语，对文本分类的作用相对来说比较显著。所以，在进行词语分类时，对于一些表现力比较弱的词，需要筛选者将其剔除，并根据其特征建立相关的集合。

在对文本进行相关分类的时候，需要通过相关统计工作，计算出字和词对应维的特征值指标，根据其特征来决定是否需要对其去除，或者使用加权操作的方法达到特征选择提取的目的，最终实现向量维数的降低及向量权数的调整。

特征抽取的算法目前有很多种， 例如， Weight of Evidence for Text(文本证据权，WET) 、Document Frequency(文档频率，DF) 、Expected Cross Entropy(期望交叉熵，ECE)等。在本系统中，我们采用词和类别的 Information Gain(信息增益，IG)，把它们作为对导入 SVM 模型的输入集特征提取的标准。

2.3 本章小结

本章首先对在线评论基础理论进行了介绍和分析，同时界定了本研究的研究范围，是针对普通用户发布的结构化汽车评论(新浪汽车网的汽车评论)和非结构化评论(新浪微博的汽车微博)；第二小节中，介绍了情感分类在情感分析领域的研究理论基础，接下来对研究过程中将涉及到的关键技术进行了详细的阐述，界定了本文的研究方法将运用有监督学习算法的方法中基于统计的机器学习算法构建情感分类模型。这一章为以下章节的实验研究提供扎实的理论基础。

第 3 章 结构化汽车评论的情感分类实验

3.1 结构化汽车评论数据的收集处理和分析

本研究的数据收集均来源于网络，通过编写程序在新浪汽车网上抓取不同时间的数据。数据的分析方法主要是统计和文本分析，即对汽车的评论文本等定性数据进行描述性统计，为导入情感分类模型预处理做铺垫。

3.1.1 结构化汽车评论样本选取

选取新浪汽车网作为采集数据并进行研究的网站主要有以下几点原因：首先，新浪汽车网是新浪网(http://www.sina.com.cn/)的汽车专版。虽然是门户网站的汽车信息网，新浪汽车网上的评论信息十分丰富，与专门的汽车网站相比毫不逊色。更重要的是，由于和中国工业汽车协会[3]的合作，新浪汽车网完整、准确地保存了大量的历史销量数据，时间跨度从 2003 年至今，可以查询到汽车销量的月度、季度和年度的销量信息。而一些专业的汽车网站上并无汽车的销量数据。第二，新浪汽车网上的汽车品牌信息和汽车评论信息十分丰富，历史数据保存完整清晰，利于做面板数据的研究。最后，新浪汽车网是国内现今最为主流的三个汽车信息网站之一，十分具有代表性。由于时间和精力的有限，本文不能将全部的三个汽车信息网站进行研究，故只采用其中之一的、比较有代表性的新浪汽车网作为研究的对象。综上所述，选取新浪汽车网的数据作为研究对象是完全合理的。

新浪汽车网上的汽车是按照车型大小进行分类的，总共有 10 类，具体包括微型车、小型车、紧凑型车、中型车、中大型车、豪华型车、多用途型车(MPV，multi-Purpose Vehicles)、运动型多用途型车(SUV，Sport Utility Vehicle)、跑车和新能源车。而根据我国最新的汽车分类国家标准(GB9417-89)[5]可以将汽车按照其用途进行分类，具体分为 7 类，即货车、越野汽车(即SUV)、自卸汽车、牵引汽车、专用汽车、客车和轿车。

为了保证收集数据的覆盖率，方便未来研究可能会对比不同类型的汽车的情感倾向性及消费者评论重点的区别，本研究将根据汽车的车型大小对样本数据进行分类，这里参照新浪汽车的分类作为分类标准，由于目前按大小、性能、用途等因素对汽车进行划分没有严格的划分标准，且用户也是根据新浪汽车的

划分进行横向比较的，因此认为这种分类是合理的，也最有利于实验设计分析。

本章研究的数据是从新浪汽车网(http://auto.sina.com.cn/)收集获得，采集程序的编写语言使用的是Java[42]，编程工具采用的是目前比较普遍的Eclipse。数据库则是利用了数据库系统MySQL，虽然要抓取的在线评论信息量大，但汽车种类是有限的，因此采用MySQL。数据采集程序分二层编程抓取数据。第一层采集汽车的基本信息参数，包括汽车车型名称，市场关注度排名，所述类型，市场价；第二层采集每个汽车车型的总评论数量，全部评论信息(包括评论星级，从 1 分到 10 分，评论文本主体，优点、缺点、综述，评论时间等)，接下来通过汽车车型这一关键词属性将两层汽车信息连接。最后，将收集的评论分成训练样本集和测试样本集，评论按评分分类，6 分到 10 分的综述评论定义为正面评论，1 分到 4 分的综述评论定义为负面评论，5 分的综述评论定义为中性评论。

本章研究从新浪汽车网上收集了所有汽车的基本信息(市场关注度排名、所属类型、市场价、上市日期等)、在线评论信息(评论星级、评论内容、评论者信息、发表时间等)。其中，汽车在线评论信息的网页截图如图 3-1。

图 3-1 汽车在线评论信息网页示例

由图 3-1 可以清晰地看出新浪汽车网允许用户使用电脑或者手机对汽车进行评论，评论的评分是用星级评判的，在采集过程中，发现网页源代码将星级

记录为”st ratings1star”到”st ratings10star”，因此评论分类时，将 1-4 的划分为负面评论，此项为 5 的是中性评论，从而 6-10 的划分为正面评论。对于用户填写的优点和缺点的划分规则是，都作为样本训练集的评论，1-4 分评论的缺点列为负面评论，优点剔除；5-10 分评论的缺点列为负面评论，优点列为正面评论。因为通过观察采集的数据发现，优点缺点栏的评论往往情感倾向性比较明显，情感词语比较丰富，因此列为样本训练集；1-4 分的用户评论往往对汽车评论比较偏激，优点栏也会有负面情感的评论出现，不具有研究价值，因此剔除此部分。因为样本量足够打，一般在分训练样本和测试样本时的比例是 7:3，因为分离出来的优点和缺点都作为训练样本，本文使用 6:4 的比例将所有评论的综述分到训练集和测试样本集中，以保证训练集和测试集符合 7:3 的比例。

3.1.2 结构化汽车评论数据整理

3.1.2.1 车型数据整理

统计新浪汽车网收录的汽车，总计共 176 个品牌，1241 款车型，885,642 条汽车评价。将在新浪汽车网上抓取全部 1241 款车型信息进行统计，可知按各车型大小分类统计分布如图 3-2 所示。统计得出，新浪新车网收录的汽车数据中车型数量最多的前三个类型为 SUV 型车，303 款占 24.42%，紧凑型车，185 款占 14.91%，跑车，143 款占 11.52%。

图 3-2 汽车类型分布百分比

首先，剔除抓取信息缺失严重的样本项，其中包括基本信息不全，或评论数为零的车型项，在参考汽车价格时，发现有很多暂时无报价的车型，无报价车型说明还没有上市，观察到评论数几乎为零，不具有参考价值，同样剔除。

剔除后得到为了在保证覆盖率和代表性的同时又可以提高分析效率，因此，我们需要找到评论数量和质量更高的车型项,去除评论量较少不具代表性的车型。我们参考评论总数属性指标进行筛选， 我们去掉了总评论数小于 10 的车型。最终得到 705 个车型,共 768,578 条在线评论,剔除后的车型类型分布如表 3-1，对筛选后的车型做了初步统计。

表 3-1 车型样本类型分布筛选前后百分比

车型	总样本	百分比	筛选后样本	百分比
微型车	69	5.56%	32	4.54%
小型车	133	10.72%	82	11.63%
紧凑型	185	14.91%	131	18.58%
中型车	128	10.31%	84	11.91%
中大型车	59	4.75%	34	4.82%
豪华型车	55	4.43%	22	3.12%
MPV	99	7.98%	54	7.66%
SUV	303	24.42%	201	28.51%
跑车	143	11.52%	55	7.80%
新能源车	67	5.40%	10	1.42%
合计	**1241**	**100%**	**705**	**100%**

我们对比筛选前后中的每种类型车所占比例可以看出，其中，新能源车所占比例变小的原因主要是新能源车都是比较新的车型，很多车型还没有上市，市场价格未定，没有用户评论，SUV 车型所占比例变大，也从某种程度侧面反映了目前市场用户的一种“SUV 热”现象。观察得到，筛选后的样本数据类型分布与总体样本集的分布大致是一样的，同样，SUV 型汽车所在比例 28.51%是最高的，紧凑型、小型车、中型车和跑车所占比例紧随其后。

接下来对评论进行剔除。

3.1.2.2 评论文本数据整理

同样，先去掉评论文字数量小于 5 个的评论项，且去掉没有评分的评论项，收集的在线评论的生成时间是从 200 年月日到 2013 年 5 月 16 日，因为数据量太大，这里对评论进行随机抽样，共 10 个类型大小的车型，针对每种类型抽取 2%的评论，得到结构化汽车评论样本 12,002 条。累计的总评论数随评论分数及发布时间的变化趋势分别如图 3-3、图 3-4 所示。

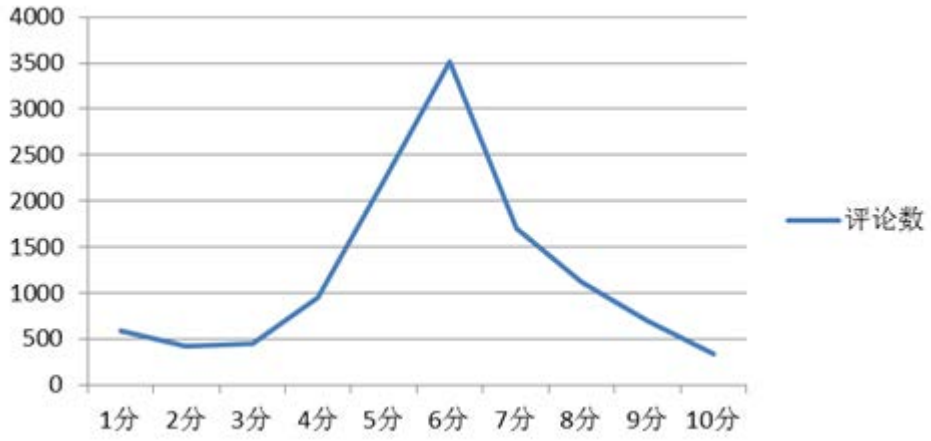


图 3-3 累计的总评论数随评论评分的变化趋势

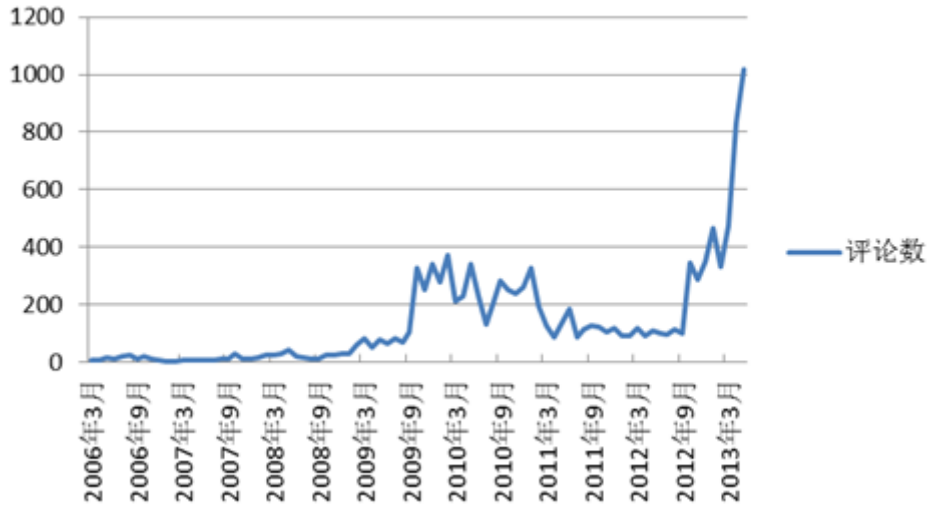


图 3-4 累计的总评论数随日期的变化趋势

因为筛选的样本是随机抽样获得，因此筛选后的样本可以反映总体汽车市场情况。由图 3-3 用户的总体评分趋势，可以清晰地看出用户对整体汽车行业的情感价值取向，用户的评分均值为 5.78。而从图 3-4 中可以看出从 2009 年汽车在线评论有了明显增加，且随着时间的推移，用户越来越喜欢在网络上发布自己对相应车型的评论。

接下来将所有在线评论的优点、缺点、综述拆分，得到了　条可用的评论数据样本，按本小节叙述的方法按照评分进行分类，最终得到训练样本集及测试样本集统计如表 3-2。

表 3-2 结构化汽车评论样本集的初步统计

	结构化汽车评论样本集					
	训练样本集			测试样本集		
	正面评论	中性评论	负面评论	正面评论	中性评论	负面评论
数量	6440	1026	4157	1348	589	1832
合计		11623			3769	
百分比		75.51%			24.49%	
总计	15392					

样本集中每条样本以一个扩展名为 txt 格式的文本形式存储在对应情感极性的文件夹中。

3.1.3 语料规范化预处理

在将结构化汽车在线评论样本集导入情感分类模型前对其进行预处理。预处理的几个关键步骤如下：

（1）分词与词性标注

对特定领域样本的分词，仅仅利用第二章中提到的中科院计算所的汉语词法分析系统(ICTCLAS3)[41]的基础训练语料库是不够的，针对我们的样本集，同时需要导入汽车行业特定专用名词和网络流行词作为用户词典。

对于网络流行词库，我们从输入法词库和搜索引擎构建的词库中各选一个导入我们的词库中。因为随着输入法技术的提高，输入法公司开始着重研究用户的拼写和输入习惯，将高频词汇，如近期出现的人或事件导入词库中，同时也会建立各领域的词库供用户自定义选择。换句话说，输入法统计出的常用词库是基于用户原创内容产生的词库，这不仅包括了传统词典中的词汇，也包括了网络热点和网络流行语句，因此，本文特选择一个覆盖率最广的输入法：搜狗输入法词库和搜狗细胞词库作为提取源之一。分词对搜索引擎的帮助同样也很大，可以帮助搜索引擎程序自动识别语句的含义，从而使搜索结果的匹配度达到最高，因此分词的质量也就直接影响了搜索结果的精确度。因此中文分词是中文搜索引擎重要的一部分，分词词库为基于词典分词的中文分词算法提供了分词的依据。经过搜索后，我们选择百度中文分词词库及相关细胞词库作为第二个提取源，最终得到两个网络流行相关词库和 11 个汽车行业细胞常用词。

为了可以达到更高的覆盖率，对于汽车领域特殊词汇的提取工作，又做了以下两部分工作。首先，先将收集到的 1141 款汽车车型导入我们的语料库中。姜亚华[45]在其研究中利用词频提取出了汽车领域的产品特征词汇表、产品评论词汇表，将这两个表的词汇也导入本研究的语库中，形成了本研究最终运用的

结构化汽车在线评论语料库。利用该语料库对目前收集的样本集进行分词工作。

利用 UltraEdit 编辑器将非文本格式的词库解析出来，统一格式后剔除重复词汇，最后得到用户词典包含网络流行高频词汇 13754 个，汽车领域评论常用词 17643 个。分为正向情感词汇 15279，负向情感词汇为 16118，导入中文分词系统将样本进行分词和标注词性处理。

（2）去除停用词

依据分词结果，接下来进行去除停用词。停用词通常是指一些应用十分广泛，但并无实质建设性意义的词，这样的词一般出现频率很高，但其作用并不能帮助提高关键词密度，反而影响主题词的准确性。停用词不同于网络流行词，针对不同的领域和使用目的，对于停用词的筛选都各有差异。在基于机器语言情感分类算法的针对汽车评论内容情感分类中，绝大多数的情感都是以形容词、动词、感叹词等修饰性词体现的，因此本文将如下几类词作为停用词删除，它们是：数词(/m，如“1”)、代词(/r，如“你”)、量词(/q，如“一辆”)、拟声词(/o，如“噢”)、方位词(/f，如“下”)、连词(/c，如“因为”)、叹词(/e，如“唉”)、后接成分(/k，如“们”)、介词(/p，如“在”)和助词(/u，如“之间”)等这些的境况。

由于这里的停用词不仅含有通用停用词，还应包括汽车领域停用词，例如，在汽车领域中，“汽车”，“轿车”一般不作为评价对象，也应作为领域停用词除去。针对不属于这几类标注的但也应当被视为停用词的词语，如“也许/d”等，本文通过参考哈工大停用词表、四川大学机器智能实验室停用词库、百度停用词列表，总结出针对微博内容的停用词表共 389 个停用词，在已有的停用词去除 JAVA 代码基础上进行改写，编译针对汽车领域的 JAVA 程序排除停用词程序。

排除停用词后，将所有评论数据项最终存为 UTF-8 格式编码的单个文本文件，此时，导入情感分类模型前的数据预处理工作准备就绪。表 3-3 是一个经预处理后的数据项的示例。

表 3-3 规范化预处理后的语料示例

示例：福克斯首选，朗逸拉皮车，速腾的定价过高，郎动就外观好看，飞翔还行。
分词后：福克斯 首选，朗逸 拉皮 车，速腾 定价 过 高，郎动 就 外观 好看，飞翔 还行。
词性标注且剔除停用词后： 福克斯/n 首选/v 朗逸/n 拉皮/v 车/n 速腾/n 定价/n 过/u 高/a 郎动/n 就/d 外观/n 好看/a 飞翔/v 还行/a

3.2 结构化汽车评论情感分类模型的建立

在本小节中，本文将详细介绍针对结构化汽车评论的情感分类模型的选取过程，模型建立的流程示意图如图 3-5 所示。分类模型的整理过程包括：数据抓取(3.1.1 小节)，语料规范化预处理(3.1.3 章节)，数据转换(特征提取、向量构造)，不同分类器的分类结果。

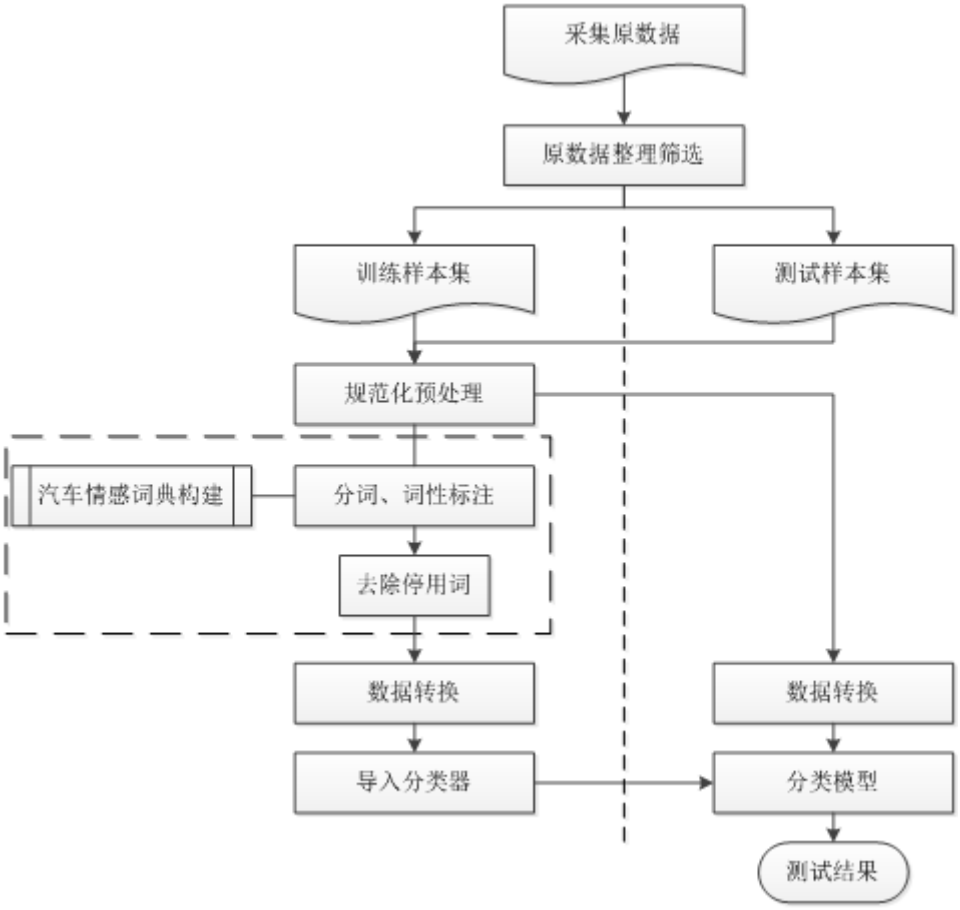


图 3-5 结构化汽车评论情感分类模型的流程示意图

3.2.1 数据转换

本文采用向量空间模型(Vector Space Model，简称VSM)进行特征提取(Term Extraction)、特征值计算并构造向量[49]。VSM 是用一些具有独立属性的字、字串、词或者短语作为项(Term)的集合成为向量空间来表示文档(Document)信息，最后通过比较向量空间之间的相似度来确定文档的归类。由于向量空间模型形式简单，又能满足一般应用的需求，传统的文本分类方法采用 VSM 均能达到

较高的准确率，因此本文选择它作为情感分类的文本表示模型。通过对数据样本进行特征提取，然后计算得到每个文本计算出的特征向量，这个最终得到形式为特征：特征权值的Index(参数列表)作为为分类器的输入数据。特征计算的流程如图 3-6 所示：

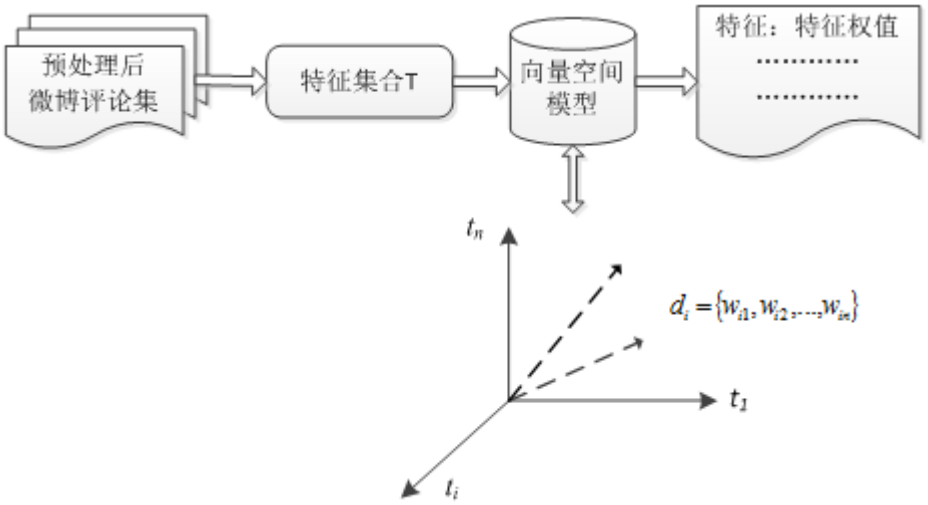


图 3-6 特征提取流程图

其中，w_{ij}为样本d_i中第j个特征项的权重，可以定义为：

$$w_{ij}=tf_{ij}*idf_i \tag{3-1}$$

tf_{ij}—— 称为词频，为向量空间中的第i个特征词在文档d_j出现的次数；

idf_i—— 为向量空间中的第i个特征词的反文档频率系数[50]。

本文特征选择方法是利用计算信息增益的方法得到的，提取的每个特征是一个分词，在 1000 维空间内对应一个特征权值。如表 3-4 所示。

表 3-4 特征值实例

序号	特征项	特征值
0	怎么	3.818711
1	笑容	6.216606
2	团结	6.216606
3	很多	4.137165
4	报告	5.523459
5	love	6.216606
6	发展	5.523459
7	模特	5.523459
8	我会	4.830312
9	得意	5.523459
…	…	…
…	…	…

得到每个分词的特征权值后，每条微博可以用一个特征向量表示，表 3-5

是一条微博的特征向量表示形式。

表 3-5 导入分类模型的文本范例(已处理好)

示例: 希望 早点 工作 买 一辆 自己 喜欢 的 车 ！奥迪 的 前脸 真心 美
Feature Vector presentation: 1 99:0.339585 310:0.359810 351:0.141443 381:0.167881 476:0.359810 477:0.282574 499:0.071576 572:0.114709

3.2.2 结构化汽车评论分类器的分类结果

如前文所述，将规范化预处理的数据样本进行数据转换后导入本文将进行对比试验，比较的三种分类器的优劣，朴素贝叶斯算法(Naïve Bayes)、支持向量机算法(SVM)、决策树 C4.5(J48)。由于本研究的主要目的是对比不同分类器的分类结果优劣，故本章后面的内容将对分类器的选取做详细的讨论。本节将对实验中将采用的结果对比进行逐一介绍，并汇报分类模型的相关实验结果。

对于结构化汽车在线评论的情感分类模型，采用 Weka 作为朴素贝叶斯网络、SVM 和决策树底层数据挖掘模型平台，利用 Java 编程语言实现了模型的训练集测试集双层训练与模型检测工作。

基于Weka平台，本研究设计并实现了结构化汽车评论情感分类算法的比较系统。Weka数据挖掘软件，全名是怀卡托智能分析环境(Waikato Environment for Knowledge Analysis)，是一款免费的、非商业化、基于Java环境下开源的机器学习软件[51]，具有非常良好的扩展性和兼容性，用户可以根据具体需要将个性化的算法封装进系统，达到数据处理及算法性能评估的目的，正是由于WEKA具有良好定义的数据结构和基本的统计接口，为自由开发者提供了一个非常有利的数据挖掘开发平台[54]。

上一小节将结构化汽车评论实验数据已经划分为训练数据集和测试数据集，WEKA 中的导入数据源文件的扩展名为.ARFF，可以用 UE 等文本编辑器打开进行改写处理。采用 Weka 进行数据分类，需要先将样本转化为.ARFF 格式的文件，因此，本小节先编写文本编译器，准备好后，开始进入数据挖掘数据分类阶段。

本章所涉及的所有实验都在一台 CPU 为 Intel Core i5 @ 2.50GHz、内存为 4.00GB、64 位 Windous 7 操作系统的 PC 上进行，Weka 版本为 Weka 3-6 稳定版，Java JDK 版本是 1.7.0。

我们将训练样本集导入分类模型进行训练，并用三种分类模型对未知数据

的测试样本集进行数据分类。实验中 Naïve Bayes 算法不需要设置参数，因为本研究的分类是线性可分的，SVM 算法的核函数采用经典的线性函数(Linear Kernel)，即

$$K(x_i, x_j) = x_i^T x_j \tag{3-2}$$

SVM 算法的参数设置如下(表 3-6)。

表 3-6 SVM 分类器的参数列表

参数类型	参数值
SVM 设置类型(SVMType)	0 -- C-SVC(默认)
损失函数(cost)	2.0
允许的终止判据(eps)	0.01
核函数类型(kernelType)	linear: u'*v
cache 内存大小(cacheSize)	40(MB)
其它(else)	缺省值

决策树 J48 算法的参数设置为“-C 0.25 –M 1”。

对分类算法的性能评估其主要依据来自于各分类算法对测试集的评估结果，根据文档测试集的实际真实类别和分类算法所分类识别出的类别之间的关系，这种表现出其之间关系的矩阵称为分类混淆矩阵。在训练时，带入预处理好的训练样本，训练分类器，再利用训练器对测试集进行分类，统计分类结果。

表 3-7、3-8、3-9 为三种分类器对测试样本分类的混淆矩阵，三个分类器的分类测试结果对比，如表 3-10 所示。

表 3-7 Naïve Bayes 分类结果的混淆矩阵

	预测 NEG	预测 NEU	预测 POS	总计
实际 NEG	1698	5	129	1832
实际 NEU	1	588	0	589
实际 POS	499	0	899	1348
总计	2198	593	1028	3769

表 3-8 SVM 分类结果的混淆矩阵

	预测 NEG	预测 NEU	预测 POS	总计
实际 NEG	1676	0	156	1832
实际 NEU	1	588	0	589
实际 POS	403	0	945	1348
总计	2080	588	1101	3769

表 3-7 决策树 C4.5(J48)分类结果的混淆矩阵

	预测 NEG	预测 NEU	预测 POS	总计
实际 NEG	1613	0	219	1832
实际 NEU	1	588	0	589
实际 POS	400	0	948	1348
总计	2014	588	1167	3769

从混淆矩阵中可以看出，对角线的数字为正确分类的样本项个数。针对三元分类，设 x=a, b, c 表示实际的类别，分类模型识别计算出来的类别为 X=A, B, C，N 为样本集的总数，则算法的精确度(Accuracy)则被定义为，

$$Accuracy = \frac{\alpha(a,A) + \alpha(b,B) + \alpha(c,C)}{N} \tag{3-3}$$

对于每个单独的类别，还有两个度量值，即召回率(Recall)和准确率(Precision)。召回率是实际为 x 类别的所有数据项中被分类算法正确归类为 x 的百分比，即

$$\mathrm{Re}\,call = \frac{\alpha(x_i, X_i)}{\sum_{j=A,B,C}\alpha(x_i, X_j)} \tag{3-4}$$

准确率(Precision)定义为确实为 x 类别的数据项占所有被分类算法归类为 x 的数据项总数的比例，即

$$\mathrm{Pr}\,ecision = \frac{\alpha(x_i, X_i)}{\sum_{j=a,b,c}\alpha(x_j, X_i)} \tag{3-5}$$

一般期望召回率和准确率同时都尽可能高，这两个指标高，则说明该分类算法对分类更有效。如果用一个指标来综合平衡这两个指标，即 F_1 测度(F_1-value，或 F Measure)，即

$$F_1 - value = \frac{2 \times precision \times recall}{precision + recall} \tag{3-6}$$

召回率(Recall)，准确率(Precision)从某一方面反应分类器的有效性。查准率(准确率)评价的是对于每一类分类情况的准确率，查全率(召回率)是不分类别的评价整个文本分类情况的准确率。全局查准率(Accuracy)等于所有分类正确的文档数与用于分类的总文档数的比值。F1 测度是一个整合指标，是一种综合了准确率与召回率的指标，更有说服力。只有当召回率和准确率的值均比较大的时候，对应的 F1 测度才比较大，它是比单一的召回率和准确率更具有代表性的指标。

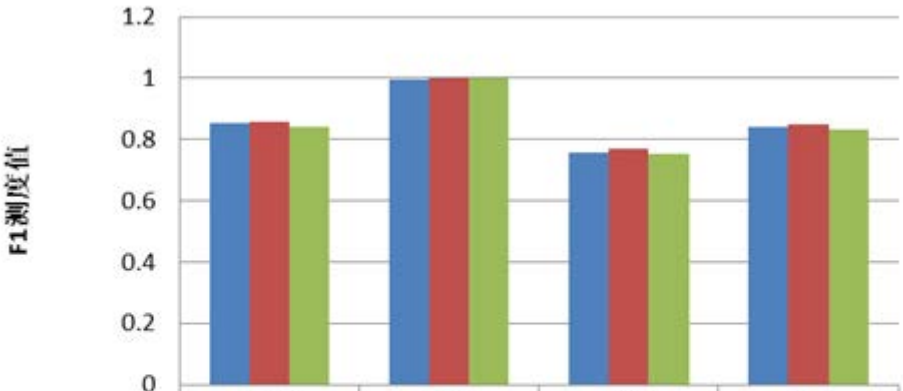
通过上述四个指标可以从多角度对模型的性能进行全面的评估。

表 3-10 三个分类器的分类测试结果对比

训练样本集数量=11623，测试样本集数量=3769						
分类器	用时(s)	精确度		召回率	准确率	F1 测度
Naïve Bayes	0.71	84.5052%	NEG	0.927	0.791	0.853
			NEU	0.998	0.992	0.995
			POS	0.667	0.875	0.757
			AVG	0.845	0.852	0.841
SVM	24.99	85.1419%	NEG	0.915	0.806	0.857
			NEU	0.998	1	0.999
			POS	0.701	0.858	0.772
			AVG	0.851	0.855	0.849
J48	219.8	83.55%	NEG	0.88	0.801	0.839
			NEU	0.998	1	0.999
			POS	0.703	0.812	0.754
			AVG	0.836	0.836	0.833

由上表可知，针对对于同一组实验数据，即针对结构化汽车评论数据，SVM 算法的分类精确度 85.14%和 F1 测度均值 0.849，均高于 Naïve Bayes 算法和决策树 C4.5 算法，但是差距不是很明显，说明三种分类模型都能较好对本研究预处理后的结构化汽车评论分类，但是看到建立训练模型的时间 J48 用时最久，Naïve Bayes 可以最快的进行分类。从实验数据可以看出每个分类器的适用情况及优缺点，Naïve Bayes 分类效率最高，但准确率相对不是很高，适合大样本数据的分类；SVM 虽然结果精确度更高，但是分类过程较繁琐，分析过程非常占用系统内存，若数据量过大，容易造成系统崩溃；J48 虽然训练时间最长，且分类结果相对逊色一些，但是可以分析含有大量噪声数据，有很好的健壮性。综上所述，针对本研究实验的样本量，采用 SVM 分类器分类效果最佳。

图 3-7 对了三个算法的 F1 测度值，从图中可以更直观看出每个算法针对每个类别识别情况，SVM 对负面评论、中性评论和正面评论的识别度都是最高的。因此，进一步证实结构化汽车评论的分类应该选用 SVM 分类器。



	NEG	NEU	POS	AVG
Naïve Bayes	0.853	0.995	0.757	0.841
SVM	0.857	0.999	0.772	0.849
J48	0.839	0.999	0.754	0.833

图 3-7 三个算法的 F1 测度值比较图

3.3 本章小结

本章首先介绍了结构化汽车在线评论数据的来源和抓取程序，接下来对收集的原数据进行筛选整理，并对预处理前的样本进行简单描述统计，首先描述了以车型为单位的数据样本，然后对以单个评论为单位的样本进行统计分析。接下来为了提高分词与词性标注工作的精准率，本研究构建了一个汽车领域情感极性词典，为了可以利用 SVM 算法进行分类，样本还利用 VSM 方法进行了特征提取，归一化处理。最后将规范化处理工作完成后的样本导入三个分类器，并对分类模型的性能进行评估工作，发现基于机器语言的分类模型都能对该样本进行有效分类，并且 SVM 模型对结构化汽车在线评论的三元分类效果最好，精确度为 85.1419%，评论 F1 测度值为 0.849。

第 4 章 非结构化汽车评论的情感分类实验

4.1 非结构化汽车评论数据的收集处理和分析

本研究第二章提到，非结构化在线评论主要指如虚拟社交网站，论坛，微博等网站用户发布的评论。据中国互联网数据平台(http://cnidp.com)，2012 年用户访问微博的频率和时间百分比远超过社交媒体网站(图 4-1)，因此，目前微博无疑是一个不可忽略的信息发布和收集平台，微博庞大的用户群体、高效传播以及高于互动性的特点使得它成为众多网站青睐的宣传阵地。因此，本章研究的数据收集均来源于非结构化在线评论的典型平台——微博。

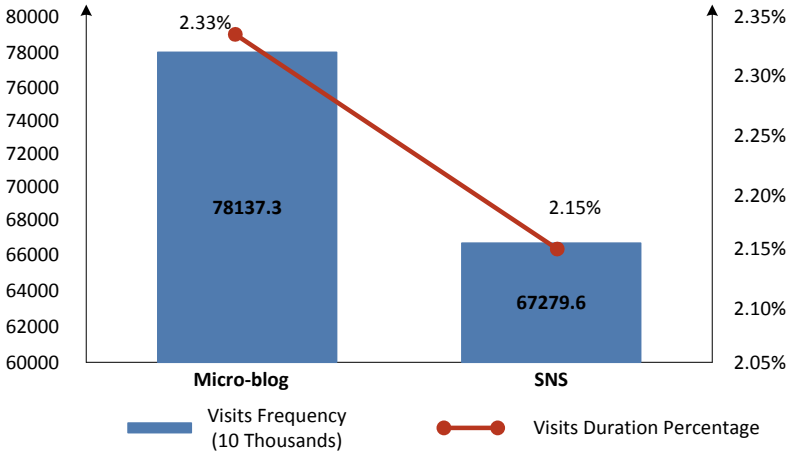


图 4-1 互联网平台用户应用的覆盖百分比

由于微博的热议和大百分比用户覆盖率，使其不仅是用户原创内容的新兴平台，同时也受到互联网研究学者的青睐。

2012 年 3 月中国互联网数据平台(CNNIC)发布的微博网站统计数据显示[2]，基于市场份额的中国微博四大平台分别为：腾讯微博、新浪微博、搜狐微博和网易微博。虽然，腾讯微博的覆盖人数最大，但通过观察可以发现，新浪微博的名人数、企业认证数量和用户活跃度及访问次数却明显超过腾讯微博，使其比腾讯微博更权威更具有用户使用行为及市场研究价值。由于本研究更注重样本的含量，对比 9 个类别关注度(新浪汽车网排名)最高的车型几个微博平台的搜索量得到图 4-2(2013 年 5 月 10 日收集数据)。

图 4-2 四大微博平台典型汽车车型的搜索微博总量

由表可知，新浪微博的微博量遥遥领先，因此，本章选择新浪微博作为微博平台研究对象，通过编写程序在新浪微博上抓取非结构化汽车评论数据。

因为本章的工作有一部分可以沿用上一章的工作成果，因此，本章将重点深入阐述在处理非结构化汽车评论额外的工作。

4.1.1 非结构化汽车评论样本选取

2012 年网站微博年度发展报告显示，2012 年在新浪微博平台中，网站官方微博仍然呈现良好乩发展趋势：截止到 2012 年底，账号数突破 4 万，与年初相比增长了 131 个百分点。根据最新数据显示，最新数据，在第三季度时，新浪微博用户数超 4 亿，而截止到 12 月底，注册用户已超 5 亿，日活跃用户为 4629 万。微博已成为网站宣传品牌、推广产品、拓展用户的重要工具。当然，汽车品牌推广也不例外，但是，本文是针对用户原创内容的研究，企业官微不在本研究的讨论范围内。新浪微博为官方机构认证有一套完善的认证机制。认证资料需要提供有效的认证申请公函以及相关的属于机构的文件证明资料。同时对于企业官方微博的头像、全称也有一定的限制。在如此严格的审查之后，官方机构微博才能够被新浪微博认证(微博名称后会有蓝 V 标记)。在数据筛选时需要抓取微博发布用户的认证级别以助于微博筛选工作。

同样，为了保证收集数据的覆盖率，本章研究数据将以新浪汽车网统计的关注度排名为参考指标，抓取关注度排名前 50 的汽车车型微博数据，也就是说，抓取以每个汽车车型作为关键字的微博。抓取非结构化汽车评论的车型列表见附录 1。

具上章统计，新浪汽车收录的汽车基本可以代表整个汽车行业，总计共 176 个品牌，1241 款车型。因为对于汽车评论爱好者来说，微博还是一个比较新兴

的平台，通过统计 2013 年第一季度(2013 年 1 月 1 日至 2013 年 3 月 31 日)关注度排名前 100 名车型的搜索原创微博(非转发)数量得到图 4-3，可以看到车型新浪汽车关注度排名和微博量无明显线性关系，但是前 50 款车型的微博量明显高于后 50 款车型的微博量，因此本章研究将收集关注度前 50 款车型的微博数据。

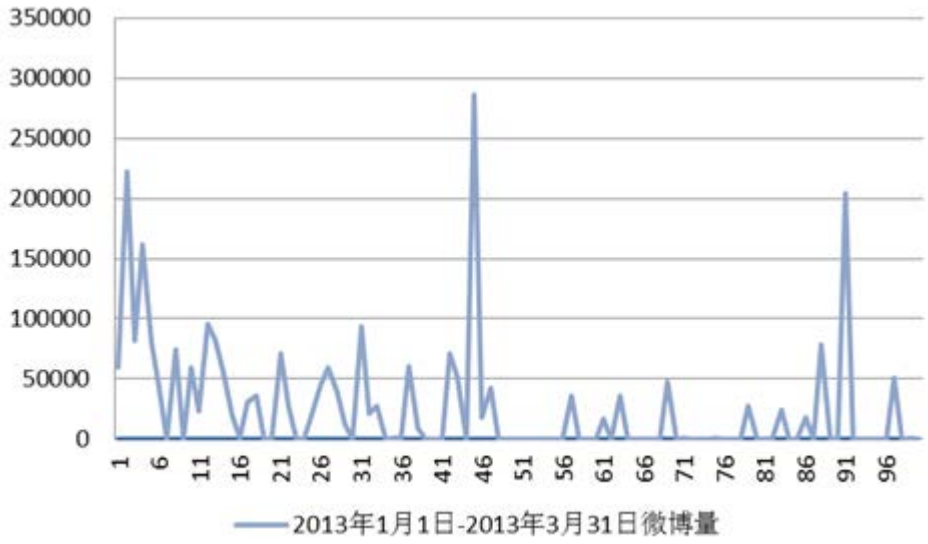


图 4-3 关注度排名前 100 名车型的搜索原创微博数量

4.1.2 非结构化汽车评论数据抓取方法简介

本节首先介绍微博数据的特征，然后介绍根据微博特征结构设计数据抓取程序。

4.1.2.1 新浪微博数据的特征

(1) 新浪微博的功能

新浪微博的功能包含评论功能，转发功能，关注功能，发布功能和搜索功能[56]。

1) 发布功能：文本限制在 140 个字内，用户可以将所想、所看、所听的信息通过 PC 或移动终端设备发布出去。即发布消息中可以包含各种多媒体，如图片、声音、视频、移动应用或产品网页链接等。

2) 转发功能：这个功能可以想象其功能是大大加速了微博消息的传播度，用户通过查看广场微博或者关注好友的微博将某条喜欢的微博进行一键转发，在转发时同时加上用户自己针对转发内容的感观体会，转发后，所有关注该用户的粉丝也能看见这条被转发消息，并且他们也可以选择继续转发，同时加入

自己的评论。

3) 关注功能：对于用户喜欢或者感兴趣的人，用户可以对其进行关注操作该被关注用户微博称为这个用户的关注者，这个用户形象地称为被关注者的粉丝。官方认证功能，加 V 用户往往是企业或名人明星，代表着更具权威性和可信度。这些人发布的微博更易引起更多用户的关注，有更大的影响力。

4) 评论功能：用户可以对任何一条微博进行评论。也可以在评论或微博消息中“@好友”，方便了快捷的反馈回执。

5) 搜索功能：根据某感兴趣的关键词搜索，可以选择搜索人或者微博。用户也可以用两个#号之间，插入某一话题进行搜索。话题性微博，不仅为民众便捷的提供了一个集中讨论的虚拟环境，而且有助于提高微博量。

(2) 新浪微博的特点

微博，微博这种非结构化在线评论与传统结构化在线评论存在着天然的联系也存在一定的差别，通过与新浪汽车网收集的结构化汽车评论进行比较，更可以凸显微博这类非结构化在线评论的特点。微博和新浪汽车网都是在线评论的发布平台，都对所有注册用户开放。虽然它们同属于短文本范畴，情感分类的粒度设置为词或短语级别有较好的分类效果，不同点在于，在直观上可以看出评论的长短上要求限制不同；其次，在评论本身结构上，两者也有所不同，微博提供用户表达感情或表述内容的形式非常丰富，可以在发布心得体会的同时，加入能更好表现心情的表情，或准确定义微博范畴可以加入话题标签(##)，可以转发，加入各种多媒体的链接，如图片、视频、声音等，微博内的关于汽车评论没有星级评分，这就需要后续的手工作业，而新浪汽车网中，口碑都是针对用户评论的车型，可以准确的定位用户的情感评论位置及情感倾向；微博的汽车评论传播范围广,用户可以随时快速的将自己对某车型感观与大家分享，新浪汽车网用户发布评论的动机与微博有差异；数据统计过程中，发现微博数据的噪声数据很大，这些噪声数据不仅仅指空数据，更多的是计算机无法自动是别的数据，如有情感表达，是个人的心情，但是用了汽车关键字在微博，而非本身针对汽车的评论，介于本章研究范畴界定的是非结构化汽车评论，首先，评论应该来源于普通用户本身，而非官方微博；其次，评论的情感应该针对于汽车本体，这对数据筛选工作是非常严峻的考验。如表 4-1 所示为两个平台的异同点。

表 4-1 微博汽车评论与新浪汽车网评论的不同点

	新浪微博评论	新浪汽车网口碑
容量	发布限制 140 字内 短文本范畴	限制在 10~500 字内 短文本范畴
结构区别	无星级评分 结构不固定，有表情、多媒体链接、转发、标签多种形式	有星级评分 结构工整，无多媒体形式，仅文字及标点符号
传播效率	及时、传播力度大	有延后性、传播范围小 决定于其他用户是否搜索
数据统计	噪声数据量很大 存在非用户原创信息 数据采集整理难度大	噪声数据少 均为用户原创信息 易于采集、整理

4.1.2.2 数据抓取程序简介

新浪为应用开发平台提供免费的API接口以供新浪微博应用程序的开发之用[57]，想要通过调用新浪微博开放平台 API 接口调用新浪微博后台数据，Consumer向 Provider 申请希望能够调用其开放 API，申请通过后由 Provider 分配给符合其要求的 Consumer ，用于唯一标识该 Consumer 符合 Provider 的要求。因此，第一步需要拥有新浪微博账号和CSDN 账号，同时用这两个账号以此获得App key 和Secret key。

因为新浪微博暂不开放微博搜索接口，利用java编程调用API仅能通过话题获取新浪微博的数据。因此，java调用API对我们要抓取的数据不可行，本章的数据抓取程序利用Gooseeker[58]公司开发的网页抓取工具MetaSeeker完成，编写java程序进行自动化数据整理。

MetaSeeker 是一个 Web 网页抓取/数据抽取/页面信息提取工具包，集成在 Firefox 浏览器的插件形式存在，能够按照用户的指导，从 Web 页面上筛选出需要的信息，并输出含有语义结构的提取结果文件(XML 文件)，众所周知，Web 页面显示的信息是给人阅读的，对于机器来说，是无结构的，MetaSeeker 解决了一个关键问题：将无结构的 Web 页面信息转换成有结构的适于机器处理的信息。可以应用于专业搜索、Mashup 和 Web 数据挖掘领域。

与普通网页不同的是，微博网站采用的编程技术使得网络爬虫抓取微博数据有很多障碍，最大的障碍是基于 Javascript/JS 的 AJAX 程序框架，导致网络爬虫很难在微博网站上爬行和抓取数据。MetaSeeker 主要包含四个工作包：

MetaStudio：是 Web 页面信息结构描述工具，提供 GUI 界面，作为 Firefox 扩展(Firefox extension)发行，推荐与 MetaCamp 和 DataStore 配套使用，这样信

息结构描述文件和各种信息提取指令文件就可以上载到 MetaCamp 和 DataStore 服务器，以拥有协同描述页面信息结构和分享信息提取成果的能力。

DataScraper：是 Web 页面信息提取(网页抓取/抽取)工具，利用 MetaStudio 生成的各种信息提取指令文件，对特定页面的信息进行连续提取，并将信息存储在 DataStore 服务器中。提供 GUI 界面，作为 Firefox 扩展发行，技术核心是一个自研的工作流引擎，由信息提取工作流指令文件驱动。

MetaCamp：是存储和管理信息结构描述文件的服务器。作为一个应用(application)部署在 Tomcat 等 Servlet 容器中。

DataStore：是存储和管理信息提取线索、各种信息提取指令文件和信息提取结果文件的服务器，集成 Lucene v2.3.2 技术，能够为结果文件建立索引。作为一个应用(application)部署在 Tomcat 等 Servlet 容器中。

因为本章研究不涉及面板数据的抓取，仅仅使用 MetaStudio 和 DataScraper 即可。同时，MetaSeeker 的部署方式具有很强的跨平台共享性，所有的 MetaStudio 定义的 Web 信息结构都将存储在云端，抓取数据时只要使用 DataScraper 从云端获取目标网页信息结构即可。

如上一节所述，每个车型有很多的评论，但是由于网页搜索仅提供 50 页的搜索量，因此针对热度排名前 50 的车型，本研究将收集将时间限定在 2013 年第一季度内网页搜索显示出的 50 页微博。

4.1.2.3 数据抓取内容设定

对微博搜索时可以只搜索原创微博，即非转发，由用户自己撰写发布的微博，高级搜索中可以通过关键词、类型、昵称、时间、地点对搜索进行筛选。如图 4-4 所示，类型：原创，时间为 2013-01-01 到 2013-03-31。

图 4-4 搜索网页筛选的设定条件

根据对新浪微博内容的分析发现，每条原创微博含有两方面信息，微博发布人的信息和微博内容。微博发布人的信息包含发布人的名字及其认证情况。微博内容包括：短链，可以通过微博跳转到其他网站、视频，亦或是 Gif 动态图片，图片内容常常与微博主题相关；话题，##的中间为话题内容，由微博发表方自行根据发送的微博内容定义主题；发布时间，即本条微博发布的时间；转发量，指微博被转述的次数；评论量，指微博被其他微博使用者评论的次数；来源，表示本条微博的发布客户端，比如新浪微博网页版，新浪微博桌面客户端，手机客户端等。

我们的研究只针对非认证用户发布的含有文本信息的微博内容本身，转发量及评论量本研究暂不考虑，因此在定义好的每个车型搜索页面中本研究只抓取三个属性：发布人名字、发布人认证情况、微博主题内容(包括话题标签及多媒体链接)。

对微博的数据抓取将使用 MetaSeeker 的 MetaStudio 模块中提供的 FreeFormat 映射结构化网页中需要采集内容所属类(class)位置，通过配置 MetaSrudio 的周期性抓取指令及相关内容，抓取以 50 个车型为关键词的指定筛选页面中每个车型搜索到的微博数量，以及列出的至多 50 页内的微博内容，每条微博内容抓取发布人名字、发布人认证情况、微博主题内容(利用 block 属性，采集指定类为 content 映射下所有的文本内容和图片，如图 4-5)，数据表采集的各字段如表 4-2 所示，每个车型的抓取规则编码见附录 2。其中车型总微博量数据在上一小节决定车型抓取量中利用。

表 4-2 数据库表的各字段参数类型描述

字段	类型	描述
keyword	文本	每个页面所属车型(主键)
count	数字	每个车型搜索出微博数
user	文本	微博发布人
centi	文本	认证状态(可以是空值)
content	文本	微博消息内容

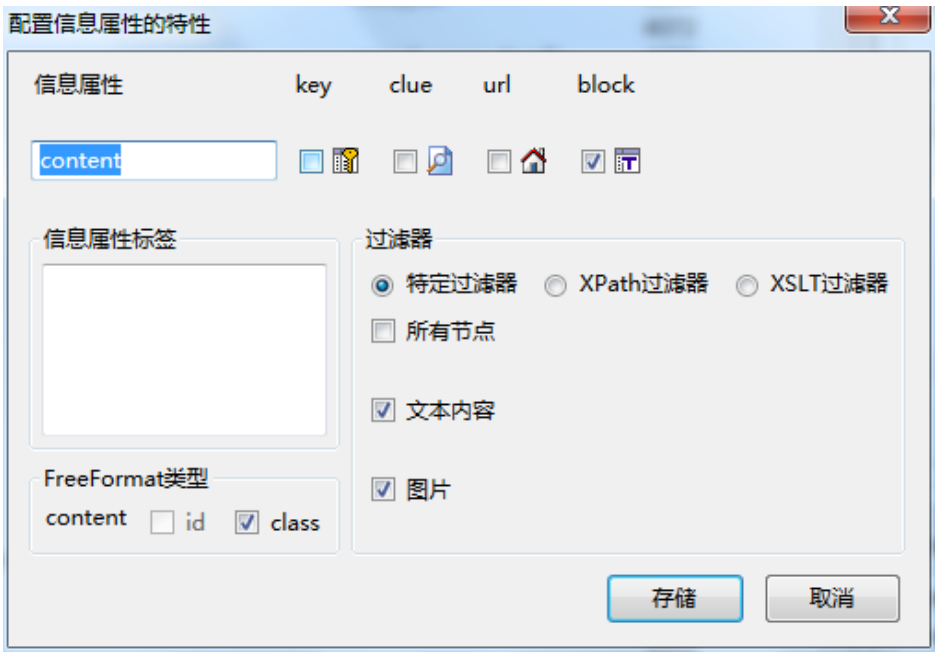


图 4-5 MetaSeeker 的配置信息属性选项

MetaSeeker 需要在 MetaStudio 中采用延长模式，在 DataScraper 提取线索时采用滚屏参数，以每 2 秒的延迟缓冲(Delay Ratio)进行滚屏操作，并在每个页面采集完成后，停留 25 秒。因为操作过于频繁程序将被新浪服务器暂时封杀，滚屏操作可以有效降低抓取空数据的概率。

4.1.2.3 XML 文件的解析整理工作

因为 MetaSeeker 输出的提取结果为 XML 文件，需要将 50 个 XML 个文件中的每条微博解析出来，分两步：第一步导入 excel 表格进行文本剔除，第二部将可用的单条微博文本形成一个单 txt 文件。

本部分利用 perl 语言对 xml 进行解析，解析成四列，分别为发布人、认证情况、微博正文(文字及标点符号，过滤@后的文字、过滤掉短链接)、微博是否含有短链接，表 4-3 中是解析出来的微博示例。

表 4-3 XML 文件解析后微博示例

user	certi	content	herf
啸舞颠狂		今天见了辆车，从前看是帕萨特……	0
盈众厦门一汽-大众同安店	新浪机构认证	一汽-大众 4 月春季大型特卖会 惠动全城-全新速腾首付 2.9 万起即可“贷”回家 ……	1
欧晓峰 edo	新浪个人认证	【全新 JEEP 自由人配 1.4T+9AT】全新 JEEP 自由人有望年内上市销售……	0
吹货 Cyt-Khuntoria	微博达人	现在看到途观第一反应就是那什么，讨厌死开途观的人了！！[怒]……	0

小宋阿然		车展归来 身体疲惫 脑袋嗡嗡的[生病] 我就纳闷了为啥年年那么吵……	0
全国经典语录大派送	微博达人	#分享视频##汽车改装#大众途观改装	1
新朗逸之家		据悉，朗逸运动版将更换采用更加运动的蜂窝状中网格栅和保险杠格栅，雾灯造型也会有变化。……	0

初步统计解析出来的微博，总计收集微博 36,942 个，发布人认证情况有六种状态：无认证、微博达人()、微博会员()、微博女郎()新浪个人认证()、新浪机构认证()，除新浪机构认证，其他状态发布人的微博均属于用户原创的在线评论范畴，新浪个人认证、微博女郎、微博达人也同时可以是微博会员；通过观察得到，有短链接的微博往往是不存在情感倾向的推广微博，这类微博对于企业决策人不具备参考价值，也不用复杂地通过情感分类模型识别，直接可以剔除。图 4-6 是收集的总微博样本中六种认证状态的发布微博数及所占百分比。

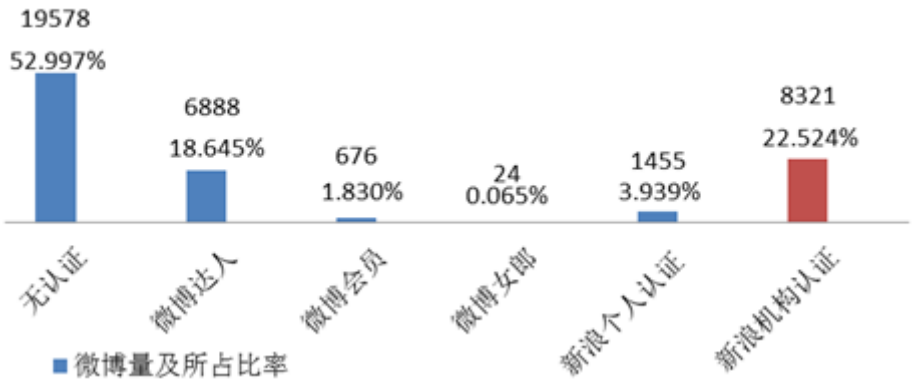


图 4-6 每类认证状态发布的微博数

对于是否存在短链接，由图 4-7 统计得到总微博样本 80.20%的微博样本无短链接，19.80%的微博附有短链接，且可以明显看出，新浪机构认证微博发布含短链接的概率比用户原创评论要高很多，说明官方微博更易发布推广类微博，因此，不在我们研究范围内。而统计得出，其他六类发布人发布的不含短链接微博数量为 23,878 条。

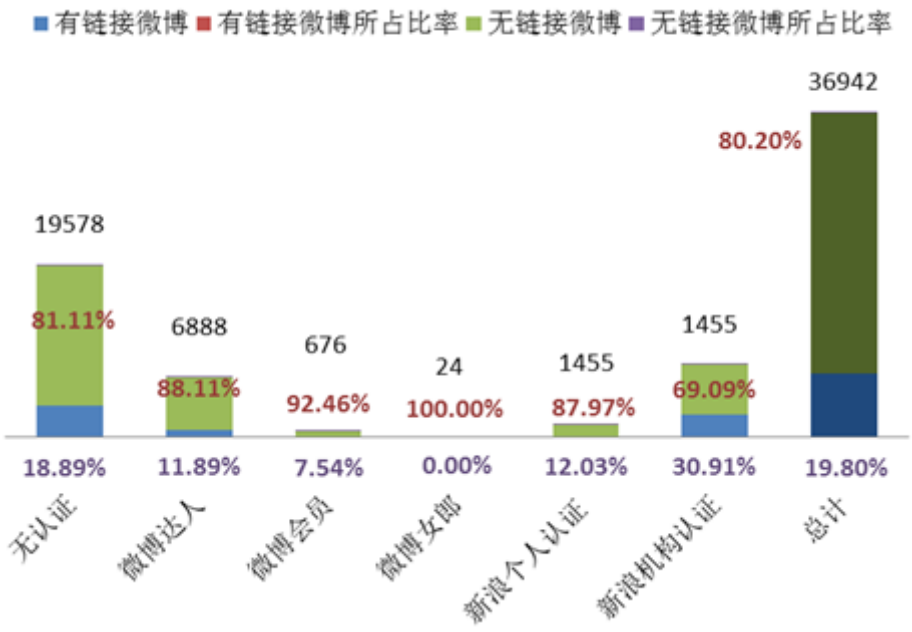


图 4-7 每类认证状态发布的是否含短链接微博统计

因此，筛选出无认证、微博达人、新浪个人认证状态发布人的无短链接微博作为研究对象，最后得到微博样本 23,878 条，将每条微博作为一个编码为 UTF-8 的 txt 文档存储。由于 23,878 条微博对于手工分类的工作量要求太大，还是利用随机抽样的方式选取了 2300 条作为分类样本。

研究邀请了哈尔滨工业大学的经济与管理学院、航天学院、计算机科学与技术学院的 4 位同学来进行用户学习(user study)，整个学习过程是匿名的，参与者通过在线问卷的形式对乐评进行手工分类。对于我们收集的汽车领域的在线评论样本，微博与结构化在线评论的区别在于，微博存在很多广告宣传、二手车买卖的信息，企业利用汽车评论情感挖掘系统需要有能力将这类微博识别出来并过滤，因此，在分类工作时，先将微博先进行主客观分类，宣传类微博定义为客观类，有情感倾向的微博定义为主观类；然后主观类微博又分为正面情感微博、负面情感微博。参与者被随机分配一些微博并将微博进行分类，如果他们认为这些候选类别都不能概括该乐评的内容，也可以通过开放式的回答来表达自己的意见，同时，每条微博都可能被一位或多位参与者手工分类，这类微博在情感上往往难以界定，剔除争议微博，最终整理得到数据统计结果如表 4-4。

表 4-4 非结构化汽车评论样本集的统计

	非结构化汽车评论样本集			
	客观情感微博样本集	主观情感微博样本集		噪声微博数据
		正面微博	负面微博	
数量		558	459	
合计	1107	1017		
百分比	52.12%	47.88%		
合计	2124			176
总计	2300			

4.1.3 非结构化汽车评论语料规范化预处理

上一小节对总体微博消除了噪声微博，换句话说，统计出来的微博都是有意义的文本内容。本小节首先对微博样本进行中文分词和词性标注。

研究过程中发现，对于微博的分词词典库，仅利用上一章改进额中科院研发的 ICTCLAS3 分词系统，还存在很多分词误差。

在统计微博数据发现，微博作为限定 140 字内的短文本，用户表达往往更碎片化，用标点符号、表情等代替了传统通顺完整的一句甚至几句话的情感表达，这使得每个词汇的信息量大大增加；且在网络平台中，文本表达以快捷、实时为初始，用户的这种随意心态使得用户的错别字率往往高于传统结构化口碑，这些无输入内容也是使得微博分词误差原因之一。这些属于噪声词汇，系统没有办法自动快速识别，且从文本内容上不足以辅助用户的语义表达。

微博还有一个不容忽视的特点，就是微博给用户提供一系列表情符号，而从上一小节中表 4-3 示例可以发现，表情作为后台数据都会被表示成转义的文字，例如[可怜]、[鼓掌]、[ok]等，因此统计微博所有提供的表情后台转义的文字表存入上一章的分词库进行改进，在对微博样本进行分词及自动词性标注。

依据分词结果，进行去除停用词工作，这里由于微博中很多标点符号也含有很大的情感信息量，不可以剔除，因此对原有停用词程序做以小修正，利用改进的微博去除停用词 Java 程序排除停用词。

同样，排除停用词后，将所有微博评论数据项最终存为 UTF-8 格式编码的单个文本文件，此时，导入情感分类模型前的数据预处理工作准备就绪。表 4-5 是一个经预处理后的数据项的示例。

表 4-5 规范化预处理后的微博语料示例

微博：昨天神秘试驾了一下 2013 款的途安，有点小失望。座椅很硬，没有电动

调节，天窗不大，门板变薄，价格不低，难道想把这个品牌做烂嘛？尼玛的[怒]
分词后：昨天/nt 神秘/a 试驾/v 了/u 一下/mq 2013/m 款/n 的/u 途安/n ，/w 有点/d 小/a 失望/v 。/w 座椅/n 很/d 硬/a ，/w 没有/v 电动/a 调节/v ，/w 天窗/n 不大/a ，/w 门板/n 变/v 薄/a ，/w 价格/n 不/d 低/a ，/w 难道/d 想/v 把/p 这个/r 品牌/n 做/v 烂/a 嘛/u ？/w 尼玛/a 的/u [怒]/a
词性标注且剔除停用词后：昨天/nt 神秘/a 试驾/v 2013/m 途安/n 有点/d 小/a 失望/v 座椅/n 很/d 硬/a 没有/v 电动/a 调节/v 天窗/n 不大/a 门板/n 变/v 薄/a 价格/n 不/d 低/a 难道/d 品牌/n 做/v 烂/a 嘛/u ？/w 尼玛/a [怒]/a

4.2 非结构化汽车评论情感分类模型的建立

在本小节中，将同样按照第三章 3.2 小节的分类模型方法，将数据导入本文重点研究的三个情感分类模型，并对比三种分类的分类结果。为了研究情感分类器对结构化评论与非结构化评论的处理差异，需要控制其他可能影响工作的一致性，因此，非结构化汽车微博情感分类模型的建立同第三章一样，采取向量空间模型进行特征提取，然后将样本转化为 ARFF 格式导入 Weka 平台进行分类测试。对于非结构化微博样本处理的不同点主要有两点：

(1) 训练集和测试集的处理方法，由于样本数量为 2124 条，相比之下样本量不是足够大，不在初始阶段做测试集和训练集的拆分，在接下来的章节中将详细阐述。

(2) 对于非结构化汽车微博评论，不单单需要对正向情感和负向情感进行区分，同样对主客观文本需要进行很好的识别，因为微博存在大量推广宣传及二手车买卖信息，这些对于企业汽车情感挖掘系统无意义。

故非结构化汽车非结构化汽车微博情感分类模型的流程如图 4-8 所示。

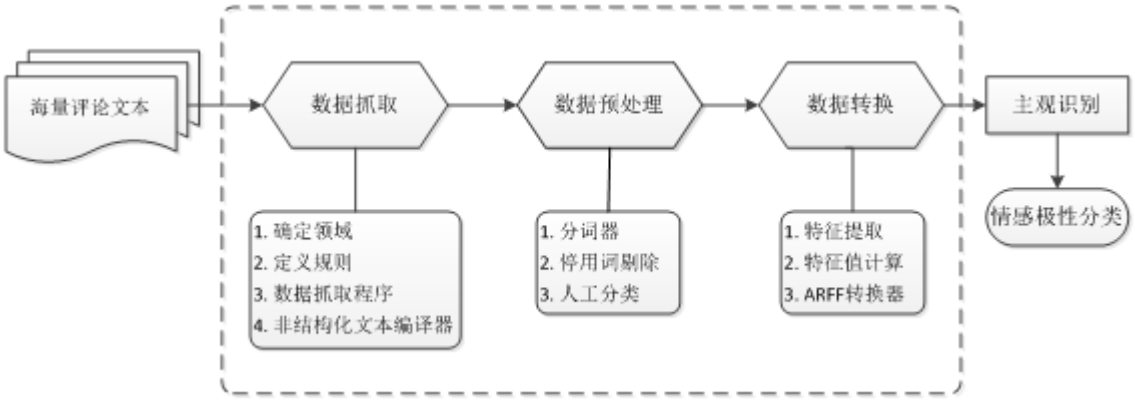


图 4-8 非结构化汽车微博情感分类模型的流程示意图

因此，对于非结构化汽车微博，包含客观微博 1107 条，正向微博 558 条，负向微博 459 条，采用三种分类器对微博进行汽车为主题的情感分类，实验包

括两个步骤：

(1) 微博的主观性内容识别；

(2) 微博的情感极性分类。

4.2.1 微博的主观性内容识别

4.2.1.1 主观性内容识别方法

张紫琼[48]等人在研究中提到，情感分析处理的文本类型是主观性文本，若要在评论情感分析系统中做到自动处理 ，首先需要区别主、客观文本信息，这是一个十分重要的处理环节，对于微博数据，存在大量的客观信息，更需要对数据进行主客观性分类。

主观性文本主要是指人们表达立场、态度和情感的想法或看法，通常带有一定的感情色彩；客观性文本指作者对事件、对象等进行基于事实的描述，不存在个人感情偏见，是客观认知的表达，推广型营销性质的微博和二手车买卖微博都是描述客观事实的微博，即也属于客观性文本的范畴。

Bharath S[59]等对 Tweets 进行情感分类研究时，首先就先运用机器学习语言分类算法对Tweets进行了主客观分类；YuH和Eric等学者分别在研究中也在篇章水平上，以词作为项，用NaiveBayes和SVM对评论与非评论进行识别，能达到相当高的分类精度[60,61]。因此针对本研究的汽车微博评论主观性内容识别，也利用Naïve Bayes、SVM、J48 尝试分类。

4.2.1.1 主观性内容识别检验评估

在研究上面提到，由于微博评论的样本量较少。在总样本集比较少时，有两种常用方法可以划分训练集和测试集以提高对模型效果进行评估检验的精确度。

(1) N 折交叉验证(N-fold cross-validation)

N 折交叉验证方法将现有数据集划分为 n 个等份相互独立的子集，利用其中一个子集作为测试集且将其余的 N-1 个子集合并作为训练集，因此一次 N-CV 的实验共需要建立 N 个模型(models)，对分类训练器进行训练，将给出 N 个精确度。那么最终训练后的分类模型的精确度为 N 个精确度的平均值。在实际操作时，N 要够大才能使各回合中的训练集样本数够多，一般而言 N=10 算是相当足够了。十折交叉验证和五折交叉验证都是比较普遍的评估方法。

(2) 留一交叉验证(LOOCV，Leave-one-out cross-validation)

假设样本集中有 n 个样本，那 LOOCV 也就是 n-CV，意思是每个样本单独

作为一次测试集，剩余 n-1 个样本则做为训练集，故一次 LOOCV 共要建立 n 个 models。相较于第一种检验方法 k-CV，LOOCV 有两个明显的优点：

一、每一回合中几乎所有的样本皆用于训练模型，因此最接近母体样本的分布，估测所得的泛化误差(generalization error)比较可靠。

二、实验过程中没有随机因素会影响实验数据，确保实验过程是可以被复制的。

但 LOOCV 的缺点则是计算成本高，因为需要建立的模型数量与总样本数量相同，当总样本数量相当多时，LOOCV 在实作上便有困难，除非每次训练模型的速度很快，或是可以用平行化计算减少计算所需的时间。

由于本研究样本集较少，对于微博分类的检验，采用十折交叉验证的方法对微博样本集进行训练并给出测试结果。

4.2.2 微博的情感极性分类

对于微博进行手工分类过程中，发现在收集的微博样本中，很少有对研究有意义的中性评价。中性情感即也是发自于用户的主观情感为前提，这类微博不属于对于客观事实的陈述，而是对个人感观的一种没有情感倾向的表达，因此也不属于客观类微博，如“淮海路偶遇天泽大众途安高配沪牌一辆”，“究竟该买途安手动还是逸致自动，大家给参考意见”。由于这类微博量很少，对这类微博的识别意义不大，不仅降低了系统效率，且占用系统资源。所以对于微博的情感极性分类只做二元分类处理，即分为正面微博和负面微博。

微博的情感极性分类工作是建立在系统先对样本进行主客观识别的基础上，因此，主客观识别中对主观情感的识别度越高，对情感极性分类精确度越高，最后得到的分类效果越好。

接下来一章将给出对于非结构化汽车微博评论的主观性内容识别结果和情感极性分类结果，并对分类器做出选择。

4.3 非结构化汽车评论分类器的分类结果

4.3.1 主观性内容识别检验结果

三种分类器分类的混淆矩阵如表 4-6、表 4-7、表 4-8，并通过三种分类器分类的精准度、召回率、准确率、F1 测度四个指标对分类器的性能进行评估，见表 4-9。

表 4-6 Naïve Bayes 分类结果的混淆矩阵

	预测 Obj	预测 Sub	总计
实际 Obj	702	405	1107
实际 Sub	24	993	1017
总计	726	1398	2124

表 4-7 SVM 分类结果的混淆矩阵

	预测 Obj	预测 Sub	总计
实际 Obj	918	189	1107
实际 Sub	102	915	1017
总计	1020	1104	2124

表 4-8 决策树 C4.5(J48)分类结果的混淆矩阵

	预测 Obj	预测 Sub	总计
实际 Obj	873	234	1107
实际 Sub	102	915	1017
总计	975	1149	2124

表 4-9 三个分类器的主客观分类测试结果对比

10 折交叉验证，Obj=1107，Sub=1017						
分类器	用时(s)	精确度		召回率	准确率	F1 测度
Naïve Bayes	0.15	79.8023%	Obj	0.634	0.967	0.766
			Sub	0.986	0.71	0.822
			Avg.	0.798	0.844	0.793
SVM	0.46	86.2994%	Obj	0.829	0.9	0.863
			Sub	0.9	0.829	0.863
			Avg.	0.863	0.866	0.863
J48	4.08	84.1808%	Obj	0.789	0.895	0.839
			Sub	0.9	0.797	0.845
			Avg.	0.842	0.848	0.842

对于主观识别的检验结果，三种分类器的精确度都在 80%左右，可以表明，利用机器语言对样本进行分类的方法可行，单从精确度一个指标看 SVM 分类器和 J48 分类器的表现略优，但是，对于其他三个指标，对于主观识别的过程

中，对客观性微博(Obj)的高识别度是毫无意义的，只有可以更精准的在数据样本中识别出主观性微博(Sub)才是我们应选择的分类器，因此，查看每个分类器对主观性微博的识别表现，如图 4-9 所示。

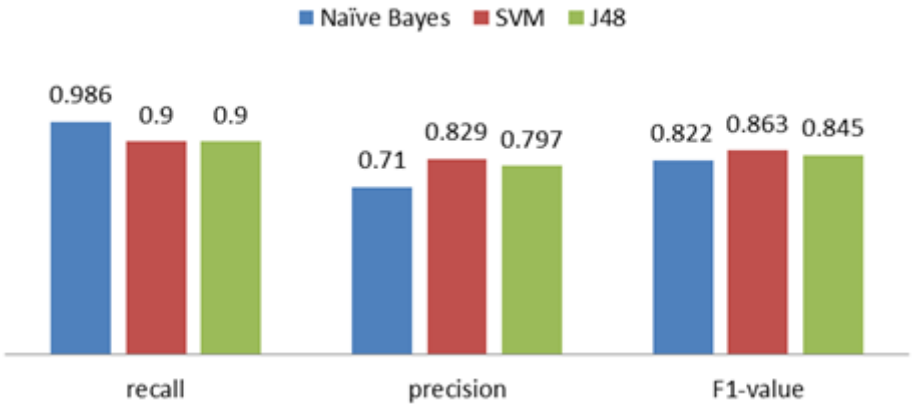


图 4-9 主观性微博(Sub)三个分类器的性能

上文提到，召回率(recall)是查看实际主观微博被正确归类为主观微博的概率，而准确率(precision)则是实际主观微博在被分类器归为主管微博的占有率，也就是被归错为主观微博的数据越少，准确率(precision)越大，越符合研究需要。由于 F1 测度是准确率和召回率的综合得分，所以对于主观识别结果，准确率的重要程度大于 F1 测度，更大于召回率。由图 4-8 可知，虽然 Naïve Bayes 的召回率最高，但是其不是最佳选择；而 SVM 分类器准确率指标值为 0.829，表现最佳。

综上所述，对于本样本的主观识别分类器仍选用 SVM 分类器。

4.3.2 情感极性分类检验结果

与主观识别方法类似，对于 558 条正面情感倾向的微博，459 条负面倾向的微博样本也进行同样的分类工作。得到的四个指标结果如表 4-10。

表 4-10 三个分类器的情感极性分类测试结果对比

10 折交叉验证，neg=459，pos=558						
分类器	用时(s)	精确度		召回率	准确率	F1 测度
Naïve Bayes	0.11	63.5659%	neg	0.972	0.609	0.749
			pos	0.211	0.857	0.338

（续表）4-10 三个分类器的情感极性分类测试结果对比

			Avg.	0.636	0.718	0.567

			neg	0.699	0.843	0.766
SVM	0.32	80.531%	pos	0.892	0.783	0.834
			Avg.	0.805	0.81	0.803
			neg	0.686	0.772	0.727
J48	3.66	76.6962%	pos	0.833	0.764	0.797
			Avg.	0.767	0.767	0.765

对于情感极性分类，研究既需要对负面评价有很好的辨识度，也需要对正面评价进行很好的识别。所以，这时与比较第三章结构化汽车评论的情感分类器参考的指标相同，先查看每个分类的精确度，负面评论(neg)、正面评论(pos)及均值的 F1 测度值也同样重要，其次是查看召回率和精确率这两个指标值。为了更直观的查看这 F1 测度的各类指标结果，给出如下图 4-9。

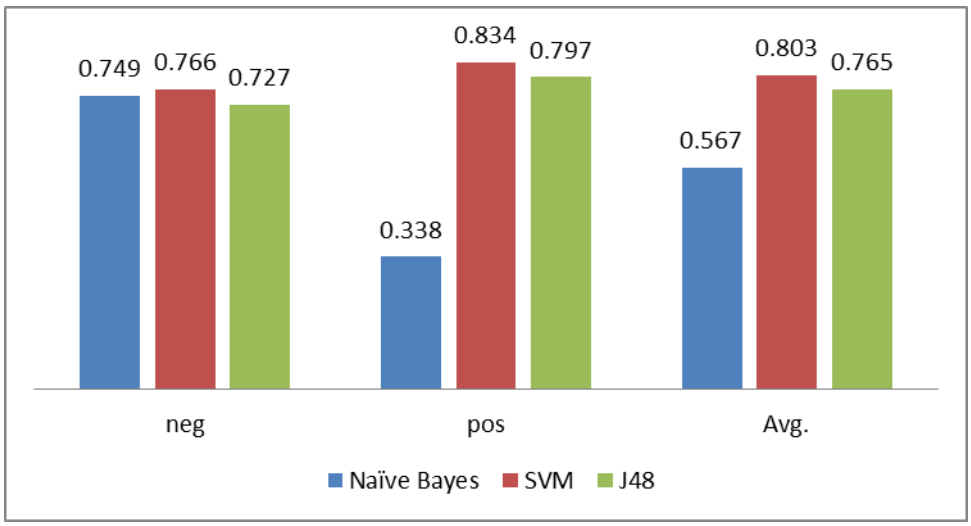


图 4-10 三个算法的 F1 测度值比较图

图 4-10 对比三个算法的 F1 测度值，从图中可以更直观看出每个算法针对每个类别识别情况，SVM 对负面评论、中性评论和正面评论的识别度都是最高的。Naïve Bayes 的精确度为 63.5659%，远低于 SVM 和 J49 的精确度，且其对正面和负面评论的辨识程度也很不平衡，放弃选用。因为本部分研究数据量较少，SVM 和 J48 分类方法均可以快速的完成训练过程。但通过第三章结构化汽车评论的处理过程可以看出，对于大样本则不然，因此，非结构化汽车评论的分类选用 SVM 分类器最佳。

4.4 本章小结

本章首先介绍了非结构化汽车评论的数据来源选取微博作为数据提取平台的原因，并分析了微博短文本的特点，比较了微博和传统结构化在线评论之间的异同。针对结构微博特有的结构，对其增加一些特殊处理，包括扩展已构建的汽车情感词典，去除含短链接的微博等，针对小样本量本部分采用十折交叉验证方法。分两步对微博进行分类，首先进行主观识别分类工作，并对分类模型的性能进行评估工作。发现基于机器语言的分类模型中，SVM 和 C4.5 决策树都能对该样本进行有效分类，而朴素贝叶斯的分类精准度只有 63.5659%，且分类稳定性较差，最后指标值可以观察得出仍是 SVM 模型对非结构化汽车在线评论的分类性能最优，精确度为 80.531%，且平均 F1 测度值为 0.803。

第 5 章 结果分析和汽车评论挖掘系统

本章将结合第 3 章结构化汽车评论的分类结果和第 4 章非结构化汽车评论的分类结果进行比较分析，建立汽车评论挖掘系统的情感分类模块。并通过构建汽车评论挖掘系统的初步框架，给出可视化界面的预构建，为汽车评论挖掘给出更直观的成果，提高本研究的企业应用价值。

5.1 汽车评论结果分析

5.1.1 分类效率的比较

首先，因为本研究对于结构化汽车评论采集了大数据样本，对非结构化汽车评论分类利用小样本数据，对大样本和小样本有一定的区别处理。大样本数据采用的随机抽样提前建立训练集和测试集，利用训练集对模型进行训练，测试集进行训练器的测试；小样本则采用十折交叉检验的方法，以保证测试集有足够的数据，保证训练器的健壮性和覆盖率。且可以很好的对比出每个分类器对大样本数据和小样本数据的处理效率，使系统保证识别度的基础上，选用效率更高的分类器。

通过我们的实验可以证实，决策树训练模型是最低效的，无论在结构化评论数据或是非结构化主客观分类及情感分类中，耗时都是最长的。Naïve Bayes 耗时最短，但是精确度也是最低的。

5.1.2 分类指标值的比较

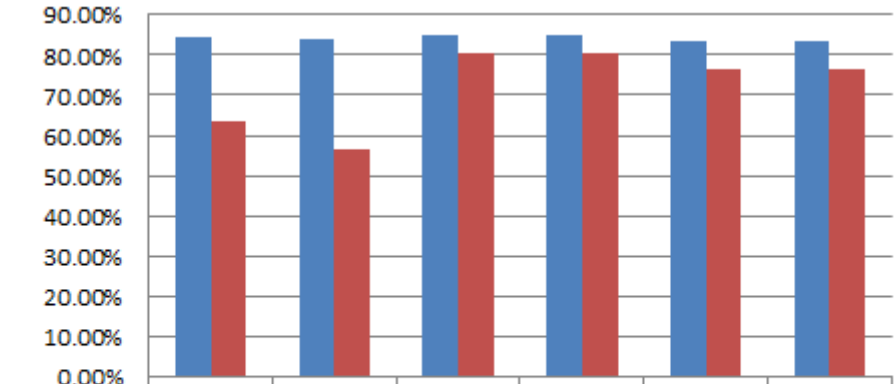
对于比较结构化分类和非结构化分类各种指标值，首先，观察比较全局性的精确度指标以及均值 F1 测度值，如图 5-1 所示。

一方面可以看出，每个蓝色柱形图都高于红色柱形图，即每个指标的第一行值都高于第二行值，也就是说，无论是这三个分类器中的哪种分类器，对结构化汽车评论分类的整体表现均优于对非结构化汽车评论的分类性能。第四章微博的特征中提到，与结构化评论相比，非结构化评论噪声数据较多，结构不工整，对预处理工作要求较高，且没有评分机制，人工分类作为标准，造成偶然误差；且样本量有所不同，样本量大，训练集中数据较多，训练器越智能，因此样本量的差距会造成一定的系统误差，从而导致每个分类器对非结构化汽

车评论的分类结果低于结构化汽车评论分类结果。

但是从非结构化微博评论的分类精确度值和 F1 测度值可以看出，可以利用机器语言对中文微博这类短文本进行分类，且选用适合样本领域的分类器可以得到较好的分类效果。对于本研究的 1017 条有情感极性的微博样本，在将汽车类评论词汇及领域特殊词汇，和微博特定流行语言、表情导入情感词典的前提下，用 SVM 分类效果，精确度达到 80.531%，平均 F1 测度值达到 0.834。因此，SVM 可以用来对中文微博进行分类。

Naïve Bayes 对非结构化评论的分类表现更要差一些。



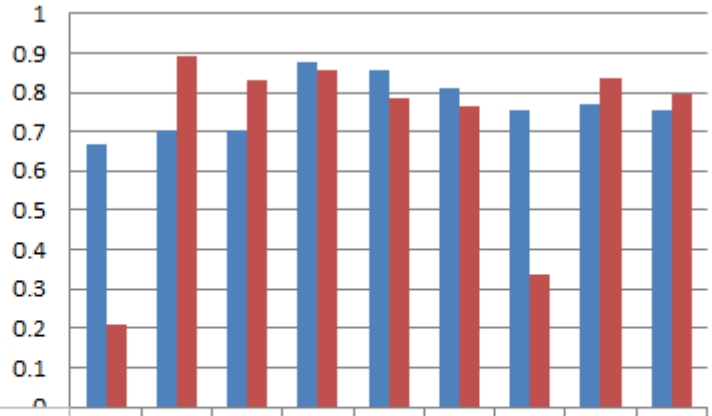
	精确度	F1测度	精确度	F1测度	精确度	F1测度
	Naïve Bayes		SVM		J48	
结构化	84.51%	0.841	85.14%	0.849	83.55%	0.833
非结构化	63.57%	0.567	80.53%	0.803	76.70%	0.765

图 5-1 精确度指标以及均值 F1 测度值

另一方面，我们的实验，对于结构化汽车评论是进行了存在中性评论的三元情感分类操作，而对于非结构化汽车评论的情感极性分类中，我们区分正面微博和负面微博，不对中性微博进行探讨。原因不是由于不存在中性微博，而是由于中性微博的特殊性，导致其对企业决策战略支持没有识别的价值。从均值 F1 测度可以看出，SVM 无论是对结构化汽车评论的三元分类，还是非结构化汽车评论的二元分类，都有很好的分类能力。

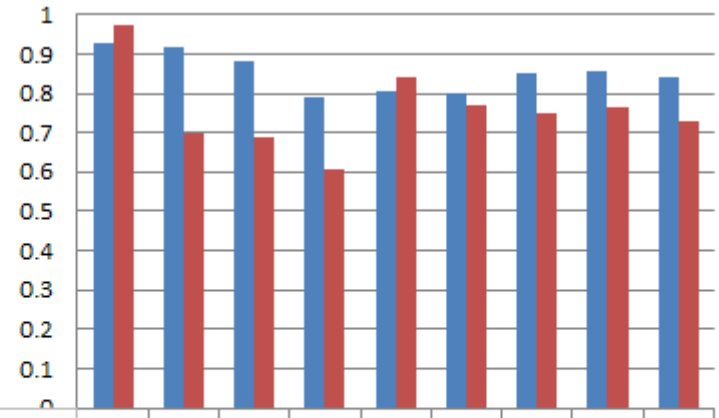
5.1.3 分类器对每类的分类结果比较

将比较三个分类对结构化评论和非结构化评论中正面评论和负面评论的识别。



Positive	NB	SVM	J48	NB	SVM	J48	NB	SVM	J48
	召回率			准确率			F1测度		
结构化汽车评论	0.667	0.701	0.703	0.875	0.858	0.812	0.757	0.772	0.754
非结构化汽车评论	0.211	0.892	0.833	0.857	0.783	0.764	0.338	0.834	0.797

图 5-2 对结构化和非结构化正面评论(pos)的分类比较



Negative	NB	SVM	J48	NB	SVM	J48	NB	SVM	J48
	召回率			准确率			F1测度		
结构化汽车评论	0.927	0.915	0.88	0.791	0.806	0.801	0.853	0.857	0.839
非结构化汽车评论	0.972	0.699	0.686	0.609	0.843	0.772	0.749	0.766	0.727

图 5-3 对结构化和非结构化负面评论(neg)的分类比较

由观察每一类(正面评论、负面评论)的分类结果，可以得出一个结论，即均值的 F1 测度能反应整体的情况，但不能代表个别类的分类性能也是一样。由上文我们得到每个分类器对结构化分类的 F1 测度都优于对非结构化汽车评论的分类性能，但是在分类器对优点进行识别时，SVM 和 J48 决策树分类器对非结构化评论分类结果高于结构化汽车评论分类结果。综上所述，选择分类器时参考正确的参数值很重要，需要整体分类效果较优时，可以参照均值 F1 测

度更为准确；但需要分类器对某类识别率更高时，需要参考个别类的 F1 测度，准确率和召回率的重要程度也是不一样的，例如第四章研究时对主观微博进行识别时就先参考主观情感分类的准确率，其次是 F1 测度和召回率值，而客观的高识别度是毫无意义的。

通过这一章节先从整体指标进行比较讨论，再进一步细化，查看每个类的三个指标值。综合考虑准确率和系统效率等因素，无论是对于结构化评论，本研究以新浪汽车网的口碑评论为例，还是对非结构化汽车评论，以微博为典型代表，利用三种情感分类器对两个数据来源进行分类，并从多维角度比较最终的分类结果，汽车评论情感挖掘系统的分类模块中最终选用 SVM 分类算法对所需样本进行自动识别分类。

5.1.4 结构化评论与非结构化评论分类的本质区别

从结果可以看出，无论是针对结构化汽车评论和非结构化汽车评论，基于机器学习的情感分类算法都可以有效对其情感进行识别，具体的实验结果如下：

(1) 基于机器学习语言的情感分类算法可以对结构汽车评论进行有效的情感分类，可以识别出正面情感、负面情感，且可以分辨出中性评论，其中 SVM 分类模型分类效果最佳。在结构化汽车评论的实验设计中，对结构化评论数据样本利用朴素贝叶斯、支持向量机、决策树 C4.5 三种情感分类器对样本进行情感极性三元分类，最终结果显示，朴素贝叶斯的分类检验指标值最低，精确度为 84.5%，平均 F1 测度值为 0.841，SVM 分类检验指标值最高，精确度为 85.14%，平均 F1 测度值为 0.849。从结果可以看出，三个分类算法的结果差异性不大，说明利用基于机器语言的情感分类模型可以对结构化汽车评论有良好的分类效果，综合精准度及训练效率等因素，对于结构化汽车评论，本研究选用 SVM 模型作为分类模型。

(2) 情感分类模型可以对非结构化汽车微博评论进行分类，不仅可以对情感极性进行较好分类，且对主客观情感可以很好的识别。在非结构化评论分类实验中，对于这部分非结构化汽车评论的分类研究分两步进行，在优异的样本主观情感识别表现的前提下，第二步进行微博的情感极性分类才具有意义。本部分研究丰富了已构建的汽车领域情感词典，使其更适用于对微博进行分词，高精确度的语料规范化处理结果是情感分类高效工作的前提。

(3) 基于机器语言的情感分类模型对汽车在线评论可以进行分类，其中 SVM 分类器，对本研究的结构化汽车评论和非结构化汽车评论样本分类都有相对最高的准确率和 F1 测度值。对于每类的 F1 测度值、召回率、准确率进行对

比观察，更能突出 SVM 分类器，分类性能稳定，表现均衡的特点。

综合以上三点研究结果得到，对于结构化汽车在线评论和非结构化汽车在线评论情感分类的本质区别在于：

(1) 分类工作流程

介于结构化在线评论和非结构化在线评论的特征，结构化在线评论结构工整，针对性强，抓取的口碑样本都是主观情感文本，不需要进行主客观情感分类；而非结构化汽车在线评论中存在大量客观文本，需要利用分类算法将主观情感进行自动化识别，因此，在分类步骤上，非结构化汽车在线评论需要先进行主观情感识别，第二步再进行微博的情感极性分类工作。

(2) 分类准确度

由于结构化在线评论，网站往往存在星级评分机制，利用表单规范引导用户进行优点缺点及自由论述，使得结构化评论易于抓取和整理，容易得到大量的训练样本，使得训练分类器的训练良好；而非结构化在线评论，需要进行人工分类等工作，导致训练样本不足够多，还存在一些系统误差和偶然误差，这些都是导致分类算法对结构化汽车在线评论表现优于非结构化汽车在线评论的原因。

5.2 汽车评论情感挖掘系统的初步构建

为提升本研究的企业应用价值，这部分将初步构建一个企业汽车评论情感挖掘系统,可以直观可视化的帮助企业先关工作人员更快的得到市场反馈信息，从海量数据中提炼有用信息，以提供决策支持与辅助。

综合利用第 3 章和第 4 章的实验研究成果，开发一个面向企业的汽车评论情感挖掘系统，本章节将对其进行概要介绍。

5.2.1 系统结构

汽车评论情感挖掘系统的目的是系统自动识别出有价值信息，可视化提供给企业相关人员进行参考，系统由前后台两部分组成。分别是：前台人机交互友好界面和后台知识库的自动生成。图 5-4 是初步的系统初步功能图。

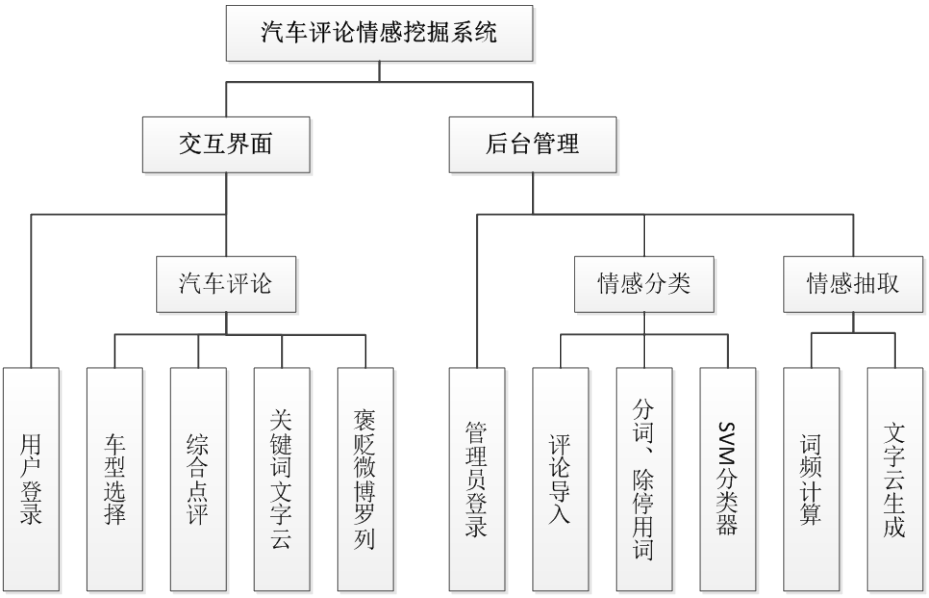


图 5-4 系统功能模块图

前台人机交互界面子系统的功能主要就是展示给企业人员关心的汽车车型的最简单最直观的市场反馈，即当用户进行登录后，选择需要的车型，系统将自动查询后台数据库。给出从结构化汽车评论出按照词频计算给出的关键字云，并罗列出以对应车型为关键字从数据库搜索出的用户褒义微博评论和贬义微博评论。

后台管理子系统，由两大主要功能模块组成，其一是本研究重点讨论的情感分类模块，其二是情感抽取模块，由于本研究的重点是对文本的情感分类研究，情感抽取模块将不再做以重点阐述对象。情感分类模块，首先，需要管理员将评论存为单个 TXT 文档以文件夹的形式导入系统数据库，点击运行调用前面研究编写的分词和词性标注程序、去除停用词程序，由于对于结构化评论和非结构化评论，前台交互界面的展现方式有所区别，所以需要选择导入文档的类别，然后运行 SVM 分类器系统自动进行分类。第二个主要功能模块，情感抽取模块，运用的关键技术就是计算每个分类结果的词频，将每个词汇和对应词频按词频排序存为 Excel 表格，第二步是将生成的 Excel 表格导入 R 语言编写的自动生成文字云程序中，存入数据库中以便前台人机交互提取查看。

5.2.2 系统交互可视化界面展示

为将情感分类结果更好的呈献给企业决策人员和市场人员，友好的可视化界面是研究应用价值的完美体现。

汽车评论情感挖掘的可视化界面，不仅可以更直观的将结果呈现，且操作

便捷，易于使用，可以减少企业人员的培训费用。

本章节将给出系统两个截屏示例：第一，前台使用界面(图 5-6)；第三，后台管理操作界面(图 5-7)。

图 5-6 前台可视化界面示例

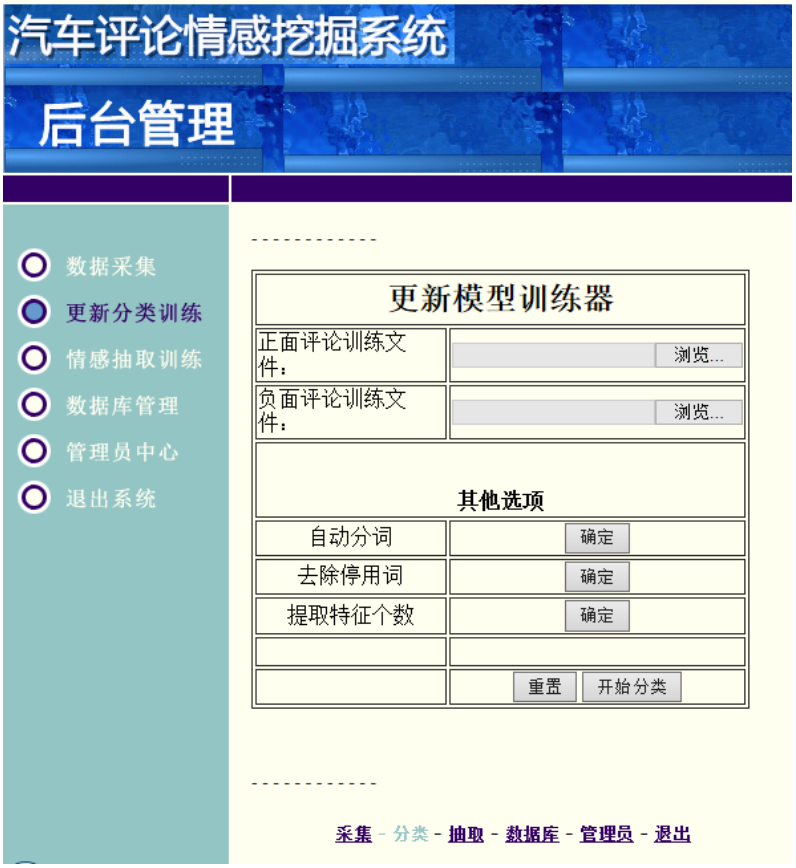


图 5-7 后台管理操作界面

这里需要阐述的是，从前台可视化界面可以看到，对速腾这款车型，系统自动将正面情感评论和负面情感评论分类，并按照词频进行情感抽取工作，这部分工作，本研究利用了第 3 章中的分词和词性标注工作，将名字+形容词，形容词+名词分成一组词，而不是全部分开，并计算词频，频率较大的词 R 语言将做字体越大的突出显示处理，词频越小则字体越小。最后得到了两篇文字云，有助于企业相关人员查看。

5.3 本章小结

本章首先通过对比研究对第 3 章结构化汽车评论和第 4 章非结构化汽车平的情感分类过程和结果进行详细的比较，并给出结果分析。将所有前文的研究工作集成于汽车评论情感挖掘系统中情感分类模块中，并建立情感抽取子系统模块，搭建初始的人机交互前台界面和后台管理维护界面，形成汽车评论情感挖掘系统雏形，为研究提升企业应用价值，为后续汽车行业在线评论的研究发展提供可行性的拓展和研究思路上的探索。

结　论

本文以汽车在线评论为研究对象，通过设计两组实验，并通过对比分析验证了基于机器学习的情感分类算法对于汽车在线评论的性能。结果显示，机器语言情感分类模型可以有效的对汽车在线评论进行分类，无论是针对结构化汽车评论和非结构化汽车评论都可以有效对情感进行识别。总结本研究的创新性工作如下：

（1） 在实验设计中，本研究将汽车在线评论分为结构化在线评论和非结构化在线评论两部分进行比较分析，介于结构化评论和非结构化评论本身的特征，对于结构化在线评论进行情感极性三元分类，而对于非结构在线评论，先进行主观情感识别，再进行情感极性二元分类研究。

（2） 在情感分类过程中，构建了汽车领域的情感极性词典，对采集样本利用编写的 java 程序进行分词、去除停用词等语料规范化处理，高精确度的语料规范化处理结果是情感分类高效工作的前提。汽车评论情感极性词典的构建提高了模型对汽车评论词汇的分辨，从而进一步提高了分类的准确率。

（3） 对于非结构化汽车在线评论的情感分类工作，研究采用微博这一研究热点平台作为数据来源，目前国内对于中文微博的情感分类研究还比较少，这使得本研究拓展了在线评论相关的研究领域，丰富了情感分类的研究范畴。且通过对比研究，分析了传统在线评论及微博类非结构化在线评论的情感分类区别。

（4） 搭建汽车评论情感挖掘系统，提升本研究的企业应用价值，为市场人员和决策人员快速、简便的提供实时的市场反馈信息及意见。系统由两个主要界面构成，人机交互前台界面和管理员后台维护界面。在前台可以可视化的展示每类车型的结构化评论情感挖掘结果和非结构化评论罗列；后台则可以对样本进行不断导入，增加数据库信息，反复训练训练器，对情感分类模块和情感抽取模块及用户登录等信息进行维护，使前台的结果更加精确。

综上所述，本文虽然没有构建新的模型，但对现有三个分类算法进行了综合比较研究，并构建了汽车评论情感挖掘系统。本研究积极探索了将在线评论情感分类方法应用到特定领域——汽车行业后，专有化改进其对于汽车在线评论的情感分类模型，并比较检验其对结构化汽车评论和非结构化汽车评论分类效果，并取得了以上的研究成果。研究之所以从结构化评论和非结构化评论两个角度出发进行研究，可以更严谨的说明机器语言对汽车在线评论的有用性，

并比较结构化评论和非结构化评论的处理差异。但由于时间、精力有限等原因，本文还存在一些不足之处有待改进：

（1） 本研究只针对实验过程中改进的汽车领域情感词典进行分词后的数据样本，逐一利用不同的情感分类模型对其进行分类，这只横向比较了三种情感分类模型对该样本的性能表现，但可以进一步研究，利用改进的汽车领域情感词典对样本数据进行规范化处理是否提高了情感分类模型的分类准确度，对于微博情感分类，可以对比没有增加汽车领域词汇、或网络流行词，以及表情文本词典，对分类效果是否存在影响。

（2） 另外，本研究是从应用性的角度出发进行研究，本着对企业相关人员的辅助进行情感分类模型研究。由于本研究采集了大量的汽车领域的数据，可以对数据进行进一步的定量分析，如评论发布时间和评论情感倾向是否存在线性关系，或者评论情感倾向对汽车销量的影响关系等。

综上所述，本研究已经取得了一定的研究成果，拥有一定的研究应用价值，但同时也存在一些不足之处。希望在未来的研究与工作中，可以完善现取得的成果，并将未完成的研究继续完成以丰富这个课题。

参考文献

[1] CNNIC发布 《第29次中国互联网络发展状况调查统计报告》[J]. 中国远程教育, 2012(02):62.

[2] CNNIC. 由中国互联网数据平台 [M]. http://www.cnnic.net.cn/.

[3] 中国汽车工业协会统计信息网 [M]. http://www.auto-stats.org.cn/default.asp

[4] 搜狐汽车网主要乘用车企业产能状况 [M]. http://auto.sohu.com/s2012/qccndc/index.shtml

[5] 易车网汽车产销统计 [M]. http://news.bitauto.com/chanxiao/

[6] Das S R, Chen M Y. Yahoo! for Amazon: Sentiment extraction from small talk on the web[J]. Management Science, 2007, 53(9): 1375-1388.

[7] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan, "Thumbs up? Sentiment Classification using Machine Learning Techniques," presented at the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP'2002), 2002: 79-86

[8] Beineke P, Hastie T, Vaithyanathan S. The sentimental factor: Improving review classification via human-provided information[C]//Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2004: 263.

[9] Morinaga S, Yamanishi K, Tateishi K, et al. Mining product reputations on the web[C]//Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2002: 341-349.

[10] Turney P D. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews[C]//Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics, 2002: 417-424.

[11] 孙文俊, 薛博召. 图书领域消费者在线评论的有用性影响因素研究[J]. 江苏商论, 2011 (5): 58-60.

[12] Wenjun S, Mingyang P, Qiang Y. Comparative Study on Objective and Subjective Emotional Tendencies of Online Reviews from Different Sources[C]//Internet Technology and Applications (iTAP), 2011 International

Conference on. IEEE, 2011: 1-4.

[13] 徐军, 丁宇新, 王晓龙. 使用机器学习方法进行新闻的情感自动分类[J]. 中文信息学报, 2007, 21(6): 95-100.

[14] 徐琳宏, 林鸿飞. 基于语义特征和本体的语篇情感计算[J]. 计算机研究与发展, 2007, 44(2): 356-360.

[15] Pak A, Paroubek P. Twitter as a Corpus for Sentiment Analysis and Opinion Mining[C]//LREC. 2010.

[16] Ethem Alpaydin. 2004. Introduction to Machine Learning (Adaptive Computation and Machine Learning). The MIT Press.

[17] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In ICML ’01: Proceedings of the Eighteenth International Conference on Machine Learning, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

[18] Barbosa L, Feng J. Robust sentiment detection on twitter from biased and noisy data[C]//Proceedings of the 23rd International Conference on Computational Linguistics: Posters. Association for Computational Linguistics, 2010: 36-44.

[19] Pang B, Lee L. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts[C]//Proceedings of the 42nd annual meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2004: 271.

[20] 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, 13(006): 84-96.

[21] 李方涛. 基于产品评论的情感分析研究[D].清华大学,2011.

[22] Liu B, Hu M, Cheng J. Opinion observer: analyzing and comparing opinions on the Web[C]//Proceedings of the 14th international conference on World Wide Web. ACM, 2005: 342-351.

[23] Schindler R M, Bickart B. Perceived helpfulness of online consumer reviews: The role of message content and style[J]. Journal of Consumer Behaviour, 2012, 11(3): 234-243.

[24] Bickart B, Schindler R M. Internet forums as influential sources of consumer information[J]. Journal of interactive marketing, 2001, 15(3): 31-40.

[25] 郝媛媛. 在线评论对消费者感知与购买行为影响的实证研究[D]. 哈尔滨工业大学, 2010.

[26] Kushal Dave, Steve lawrence, and David M.Pennock. 2003. Mining the peanut gallery Opinion extraction and semantic classification of product reviews. In Proceedings of the international World Wide Web conference[C]. 519-528.

[27] Baccianella S, Esuli A, Sebastiani F. SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining[C]//LREC. 2010, 10: 2200-2204.

[28] Heyer G, Quasthoff U, Wittig T. Text Mining: Wissensrohstoff Text[M]. W3L-Verlag, 2006.

[29] Fellbaum C. WordNet[M]. Springer Netherlands, 2010.

[30] Ohana B, Tierney B. Sentiment classification of reviews using SentiWordNet[C]//9th. IT & T Conference. 2009: 13.

[31] Pang B, Lee L. Opinion mining and sentiment analysis[J]. Foundations and trends in information retrieval, 2008, 2(1-2): 1-135.

[32] Kim S M, Hovy E. Automatic identification of pro and con reasons in online reviews[C]//Proceedings of the COLING/ACL on Main conference poster sessions. Association for Computational Linguistics, 2006: 483-490.

[33] Baccianella S, Esuli A, Sebastiani F. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining[C]//Proceedings of the 7th conference on International Language Resources and Evaluation (LREC'10), Valletta, Malta, May. 2010.

[34] 董婷. 支持向量机分类算法在 MATLAB 环境下的实现[J] . 榆林学院学报, 2008, 18(4) : 94- 96.

[35] LI J M, Zhang B, LIN F Z. An improvement algorithm to sequential minimal optimization[J]. Journal of Software, 2003, 5: 007.

[36] Platt J. Sequential minimal optimization: A fast algorithm for training support vector machines[J]. 1998.

[37] Hunt D E, Joyce B R. Teacher trainee personality and initial teaching style[J]. American Educational Research Journal, 1967: 253-259.

[38] Quinlan J R. Induction of decision trees[J]. Machine learning, 1986, 1(1): 81-106.

[39] Quinlan J R. C4. 5: programs for machine learning[M]. Morgan kaufmann, 1993.

[40] 于士涛. 基于问答网络论坛知识体系的自动问答系统研究[D]. 南开大学, 2009.

[41] 汉语词法分析系统ICTCLAS. http://www.ict.ac.cn/jszy/jsxk_zlxk/mfxk/20-0706/t20070628_2121143.html

[42] Java. http://www.java.com/

[43] Zhang H P, Liu Q. Automatic Recognition of Chinese Personal Name Based on Role Tagging[J]. CHINESE JOURNAL OF COMPUTERS-CHINESE EDITION-, 2004, 27(1): 85-91.

[44] An J X, Huang J, Yu W. Algorithm of Disambiguation and Matching of Chinese Word Segmentation in Connected Strategies Research[J]. Advanced Materials Research, 2011, 219: 1702-1706.

[45] 姜亚华. 基于 Hownet 的汽车领域产品评论挖掘方法研究[D]. 哈尔滨工业大学, 2011.

[46] 游建平. 基于语义情感空间模型的微博情感倾向性研究[D].暨南大学,2012.

[47] Gamon M, Aue A. Automatic Identification of Sentiment Vocabulary: Exploiting Low Association with Known Sentiment Terms[C]//Proc. of the ACL-2005 Workshop on Feature Engineering for Machine Learning in NLP. Michigan, USA, 2005: 57-64.

[48] Eric T.G Wang,Paul Jen-Hwa Hu. EXAMINING THE ROLE OF INFORMATION TECHNOLOGY IN CULTIVATING FIRMS' DYNAMIC MARKETING CAPABILITIES[A]. Department of Decision Sciences and Managerial Economics、The Chinese University of Hong Kong.Proceedings of the Ninth International Conference on Electronic Business[C].Department of Decision Sciences and Managerial Economics、The Chinese University of Hong Kong:,2009:11.

[49] Radev D R, Jing H, Styś M, et al. Centroid-based summarization of multiple documents[J]. Information Processing & Management, 2004, 40(6): 919-938.

[50] Aizawa A. An information-theoretic perspective of tf–idf measures[J]. Information Processing & Management, 2003, 39(1): 45-65.

[51] Weka. http://www.cs.waikato.ac.nz/ml/weka/

[52] 潘明暘. 不同来源在线评论对消费者行为影响研究[D].哈尔滨工业大学,2011.

[53] Yu H,Hatzivassiloglou V.Towards answering opinion questions: separating facts from opinions and identifying the polarity of opinion sentences[C]// Proc. of the 2003 Conf. on Empirical Methods in Natural LanguageProcessing. Sapporo, Japan, 2003: 129-136.

[54] Witten I H, Frank E. Data Mining: Practical machine learning tools and techniques[M]. Morgan Kaufmann, 2005.

[55] Witten I H, Frank E, Hall M A. Data Mining: Practical Machine Learning Tools and Techniques: Practical Machine Learning Tools and Techniques[M]. Morgan Kaufmann, 2011.

[56] 新浪微博. Sina Weibo. http://www.weibo.com.

[57] 新浪微博API. Sina Weibo API. http://open.weibo.com/wiki/%E5%BE%AE-%E5%8D%9AAPI

[58] Gooseeker. http://www.gooseeker.com/.

[59] Sriram B, Fuhry D, Demir E, et al. Short text classification in twitter to improve information filtering[C]//Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval. ACM, 2010: 841-842.

附录 1 非结构化汽车评论抓取车型

关注度排名	车型	类型	类型排名	新浪评论数
1	速腾	紧凑型	9	7003
2	朗逸	紧凑型	8	4536
3	迈腾	中型	5	5002
4	福克斯	紧凑型	1	11102
5	途观	SUV	3	2240
6	凯越	紧凑型	11	15903
7	奥迪 A4L	中型	7	2253
8	凯美瑞	中型	10	4355
9	SX4	紧凑型	30	2089
10	雅阁	中型	14	16960
11	CR-V	SUV	12	3706
12	明锐	紧凑型	19	5648
13	捷达	紧凑型	6	36583
14	科鲁兹	紧凑型	12	4001
15	新宝来	紧凑型	3	2328
16	飞度	小型	12	16676
17	卡罗拉	紧凑型	21	2724
18	马自达 6	中型	16	14355
19	逍客	SUV	18	2241
20	奥迪 A6L	中大型	2	6313
21	帕萨特	中型	4	14860
22	马自达3经典款	紧凑型	47	2037
23	轩逸	紧凑型	15	2928
24	骊威	小型	8	1837
25	天籁	中型	18	4558
26	君威	中型	11	7967
27	高尔夫	紧凑型	10	6063
28	思域	紧凑型	33	3653

29	赛欧	小型	6	3783
30	标致 307	紧凑型	49	8303
31	奥迪 Q5	SUV	7	780
32	君越	中型	15	3199
33	途安	MPV	3	2615
34	伊兰特	紧凑型	29	10876
35	骐达	紧凑型	25	3373
36	锋范	小型	14	2024
37	宝马 5 系	中大型	3	1339
38	汉兰达	SUV	14	1255
39	B70	中型	41	2251
40	B50	紧凑型	46	2287
41	蒙迪欧致胜	中型	8	8977
42	RAV4	SUV	25	1039
43	宝马 3 系	中型	9	1726
44	比亚迪 F0	微型	2	1723
45	桑塔纳	紧凑型	7	4205
46	雨燕	小型	16	6487
47	大众 CC	中型	6	1264
48	圣达菲	SUV	83	964
49	骏捷 FSV	紧凑型	22	845
50	C5	中型	26	3289
51	骏捷	中型	48	8580
52	途胜	SUV	55	1543
53	威驰	小型	27	10663
54	狮跑	SUV	56	746
55	悦动	紧凑型	24	3194
56	花冠	紧凑型	34	6474
57	锐志	中型	22	3032
58	哈弗 H3	SUV	76	1696
59	马自达 2	小型	15	739
60	新奥拓	微型	1	2053

61	爱丽舍	紧凑型	63	16875
62	350	紧凑型	5	1987
63	途锐	SUV	23	273
64	东风标致 408	紧凑型	26	2545
65	普拉多	SUV	45	803
66	乐驰	微型	5	2900
67	瑞虎	SUV	50	3619
68	ix35	SUV	11	697
69	Polo	小型	4	6059
70	赛拉图	中型	43	4154
71	景程	中型	35	3274
72	奔驰 E 级	中大型	1	850
73	雅绅特	小型	20	2681
74	新皇冠	中大型	5	1783
75	森林人	SUV	39	583
76	骏捷 CROSS	紧凑型	111	425
77	比亚迪 F6	中型	52	998
78	宏光	MPV	1	1416
79	奥迪 Q7	SUV	1	382
80	英朗	紧凑型	13	1790
81	世嘉	紧凑型		
82	晶锐	小型	18	1641
83	奥德赛	MPV	5	1612
84	QQ	微型	7	8324
85	夏利 A+	小型	28	3059
86	尊驰	中型	69	3011
87	昊锐	中型	25	1509
88	帝豪 EC7	紧凑型	20	2279
89	卡宴	SUV	32	214
90	MG3	小型	1	1300
91	福瑞迪	紧凑型	64	1358
92	奔驰 C 级	中型	17	569

93	锐欧	小型	37	948
94	兰德酷路泽	SUV	53	192
95	MG6	紧凑型	44	1683
96	雅力士	小型	30	751
97	奇瑞 A3	紧凑型	79	4575
98	宝马 7 系	豪华型	3	306
99	睿翼	中型	32	1443
100	神行者	SUV	16	166

附录 2 非结构化汽车评论网络爬虫

```
<?xml version="1.0" encoding="UTF-8"?>
<xsl:stylesheet version="1.0"
xmlns:xsl="http://www.w3.org/1999/XSL/Transform">
<xsl:template match="/">
<page>
<xsl:apply-templates select="//*[@id='pl_common_searchTop' and
count(.//*[@class='search_input']/div/div[position()=2]/div/input/@value)>0]"
mode="page"/>
</page>
</xsl:template>
<xsl:template name="weibo_">
<item>
<user>
<xsl:value-of
select="*//*[@class='content']/p[position()=1]/a[position()=1]/@title"/>
<xsl:value-of
select="*[@class='content']/p[position()=1]/a[position()=1]/@title"/>
<xsl:if test="@class='content'">
<xsl:value-of select="p[position()=1]/a[position()=1]/@title"/>
</xsl:if>
</user>
<certi>
<xsl:value-of
select="*//*[@class='content']/p[position()=1]/a[position()=2]/img/@alt"/>
<xsl:value-of
select="*[@class='content']/p[position()=1]/a[position()=2]/img/@alt"/>
<xsl:if test="@class='content'">
<xsl:value-of select="p[position()=1]/a[position()=2]/img/@alt"/>
</xsl:if>
</certi>
<content>
<xsl:value-of select="*//*[@class='content']"/>
<xsl:value-of select="*[@class='content']"/>
```

```
</content>
</item>
</xsl:template>
<xsl:template match="//*[@id='pl_common_searchTop' and
count(.//*[@class='search_input']/div/div[position()=2]/div/input/@value)>0]"
mode="page">
<item>
<keyword>
<xsl:value-of
select="*//*[@class='search_input']/div/div[position()=2]/div/input/@value"/>
<xsl:value-of
select="*[@class='search_input']/div/div[position()=2]/div/input/@value"/>
<xsl:if test="@class='search_input'">
<xsl:value-of select="div/div[position()=2]/div/input/@value"/>
</xsl:if>
</keyword>
<count>
<xsl:value-of
select="following-sibling::div[position()=1]/div/div[position()=1]/div/span/text()"/
>
</count>
<weibo_>
<xsl:for-each
select="following-sibling::div[position()=1]//*[@class='feed_lists W_linka
W_texta']/dl[position()>=1]">
<xsl:call-template name="weibo_"/>
</xsl:for-each>
</weibo_>
</item>
</xsl:template>
</xsl:stylesheet>
```

附录 3 XML 数据解析部分程序

```
use XML::Simple;
use 5.014;

my $dir_name = "D:/Research/SinaWeibo/";
opendir( DIR, $dir_name ) || die "Can't open directory $dir_name";
my @dots = readdir(DIR);
foreach (@dots) {
    if ( index( $_, "-" ) > -1 ) {
        my $output_folder = $_;
        my %tweets        = ();
        print mkdir ("tweet/day".$output_folder, 0777);#make directory for each day

        my $sub_dir = $dir_name . $_ . "/";
        opendir( SUB_DIR, $sub_dir ) || die "Can't open directory $sub_dir";
        my @files = readdir(SUB_DIR);

        #parse weibos and remove duplicate weibos by hash
        foreach (@files) {
            if ( index( $_, ".xml" ) > 0 ) {
                print $_."\n";
                my @file_props = stat( $sub_dir . $_ );
                if ( $file_props[7] > 20480 ) {
                    my $simple = XML::Simple->new( KeyAttr => "id" );
                    my $data = $simple->XMLin( $sub_dir . $_ );
                    my %current_tweets = %{ $data->{status} };
                    foreach my $key ( keys(%current_tweets) ) {
                        #my $text = $current_tweets{$key}{"text"};
                        my $content =
$key."|".$current_tweets{$key}{"user"}{"followers_count"}."|".$current_tweets{$key}{"c
reated_at"};
```

```
                                #$tweets{$key} = $text;
                                $tweets{$key} = $content;
                            }
                        }
                    }
                }

    #write weibos to a txt file
                open (LOG, '>:encoding(UTF-8)',"map/".$output_folder.".txt") or die "Can't
open\n";
                foreach my $key(keys %tweets){
                    print $key."\n";
                    print LOG $tweets{$key}."\n";
                }
                close LOG;
                #write each weibo to a single txt file
    #           foreach my $key (keys %tweets){
    #               open (BEDROCK,
'>:encoding(UTF-8)',"tweet/day".$output_folder."/".$key.".txt") or die "Can't open!\n";
    #               print "tweet/day".$output_folder."/".$key.".txt"."\n";
    #               print BEDROCK $tweets{$key}."\n";
    #               close(BEDROCK);
    #           }

                print $_. "-------------------------------------------------\n";
            }
    }
    closedir DIR;
    print "finish";
```

攻读硕士学位期间发表的论文及其它成果

He Huang. Sentiment Analysis of Sina Weibo based on Semantic Sentiment Space Model. ICMSE 2013[C]//20th Annual Conference Proceedings, International Conference on Volume, Issue. Accepted. /中文：黄鹤. 基于语义情感空间模型的中文微博情感分析, 2013 年管理科学与工程国际会议(第 20 届)，已录用

哈尔滨工业大学学位论文原创性声明和使用权限

学位论文原创性声明

本人郑重声明：此处所提交的学位论文《面向汽车在线评论的情感分类研究与应用》，是本人在导师指导下，在哈尔滨工业大学攻读学位期间独立进行研究工作所取得的成果，且学位论文中除已标注引用文献的部分外不包含他人完成或已发表的研究成果。对本学位论文的研究工作做出重要贡献的个人和集体，均已在文中以明确方式注明。

作者签名：  日期：2013 年 7 月 3 日

学位论文使用权限

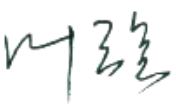
学位论文是研究生在哈尔滨工业大学攻读学位期间完成的成果，知识产权归属哈尔滨工业大学。学位论文的使用权限如下：

（1）学校可以采用影印、缩印或其他复制手段保存研究生上交的学位论文，并向国家图书馆报送学位论文；（2）学校可以将学位论文部分或全部内容编入有关数据库进行检索和提供相应阅览服务；（3）研究生毕业后发表与此学位论文研究成果相关的学术论文和其他成果时，应征得导师同意，且第一署名单位为哈尔滨工业大学。

保密论文在保密期内遵守有关保密规定，解密后适用于此使用权限规定。

本人知悉学位论文的使用权限，并将遵守有关规定。

作者签名：  日期：2013 年 7 月 3 日

导师签名：  日期：2013 年 7 月 3 日

致　谢

光阴似箭，四年工大的本科学习和两年的研究生学习生活即将结束。回首这留念的大学光阴，对那些引导我、帮助我、激励我的人，我心中充满了感激，这个校园，也让即将离校工作的我，恋恋不舍。

本论文是在我的导师叶强教授的悉心指导下完成的。在这近一年的毕业论文研究期间，叶老师渊博的专业知识、严谨的治学态度、谦虚的品质和儒雅的为人谈吐，给我留下了深刻的印象。他精益求精的工作作风，和谦逊温和的人格魅力对我以后的做人做事影响深远。叶老师作为我本科四年，研究生两年的导师，他使我不仅掌握了许多知识和做学问的方法，也学到了不少做人的道理。衷心感谢叶老师的耐心指导，亲切教诲。

感谢方斌师兄的认真指导和帮助。毕业论文期间在编程技术和数据处理知识上方斌师兄给我很多的帮助，使我在研究中有了非常大的灵感。同时感谢我的室友，郭珊珊、张文莹、向蕊沁，在我论文分类工作上给予我论文的帮助。

感谢我的朋友司胜营，在他的帮助下，我了解了很多汽车行业的专业知识和领域，为本篇论文的顺利完成，和我即将工作涉及的工作领域，做了很多准备工作，使我更有自信，更有能量。

感谢同在哈尔滨工业大学共同完成毕业设计的所有同学，和他们度过的这些快乐时光，是我非常美好的回忆。

最后，我要诚挚感谢我最亲爱的父母在生活和学习中对我的鼓励和关怀。正是在他们的支持和理解下，我可以快乐学习，无忧的生活，本论文才得以顺利完成。

即将完成两年的硕士生涯，迈入人生的另一个阶段，心中感慨万千。感谢工大给我的整整六年绚丽的大学校园生活，未来的日子里我要继续发扬工大的研究精神，做一个合格的工大人！