硕士学位论文

基于互联网评论的股票市场趋势预测

STOCk MARKET TREND FORECAST BASED ON INTERNET REVIEWS

李玉梅

哈尔滨工业大学

2012 年 6 月

国内图书分类号：F830.91　　学校代码：10213

国际图书分类号：336　　密级：公开

管理学硕士学位论文

基于互联网评论的股票市场趋势预测

硕士研究生：李玉梅

导　　师：闫相斌

申请学位：管理学硕士

学　　科：管理科学与工程

所在单位：管理学院

答辩日期：2012 年 6 月

授予学位单位：哈尔滨工业大学

Classified Index: F830.91

U.D.C: 336

Dissertation for the Master Degree in Management

STOCk MARKET TREND FORECAST BASED ON INTERNET REVIEWS

Candidate:	Li yumei
Supervisor:	Prof. Xiangbin Yan
Academic Degree Applied for:	Master of Management
Speciality:	Management science and Engineering
Affiliation:	School of Management
Date of Defence:	June, 2012
Degree-Conferring-Institution:	Harbin Institute of Technology

摘　要

随着网络在国内普及率飞速增长，网络信息量呈几何级数增长，其传播的速度更是其它渠道难以匹敌的， 成为人们最重要的信息来源之一。网络也成为金融领域信息重要的集散地，尤其是 WEB2.0 技术的发展，论坛、博客、聊天室等可以提供互动的技术不断涌现，使投资者可以参与到网络信息的创造、传播及获取的各个环节。论坛是最受欢迎的网络社区之一，众多的投资者在股票论坛中交流信息，分享经验以辅助投资决策，因此对其中信息的获取是了解投资者心理及行为的重要途径。

相比国外上百年历史的成熟的金融市场，成立仅二十余年的中国金融市场还处于发展阶段，监管制度不完善，投机者居多。众多投资者通过各种途径获取信息进行交易，作为获取信息的重要方式之一，对股票论坛的研究具有重要意义。

行为金融理论认为投资者的心理及行为能够影响股票市场的表现，基于这一理论，本文对国内的股票市场进行了研究。本文提出了自动剔除领域无关评论的方法，成功剔除了 84%的股票市场无关评论，并保留了 90%以上的股票市场相关信息。本文对比了语义分析方法、机器学习方法及 N-Gram 方法三种情感分析方法，支持向量机结合信息增益的方法能够获得良好的实验结果。通过单只股票价格影响因素分析，建立股票价格预测模型，能够比较准确地预测股票市场的价格。

我们分析了股票价格影响因素，并建立回归模型对其进行预测。结果显示，滞后股票收盘价，情感指数，机构评分、滞后新闻数量能投对股票收盘价格进行解释。通过对通讯行业进行单因素方差分析，情感指数能够影响收益率及波动率。通过对上证指数及情感指数进行领先滞后分析发现，投资者情绪与滞后综合指数相关，与领先个股收盘价相关。

关键词：股票市场；股票评论；情感分析；股票价格；预测

Abstract

As the network gets a rapidly popularity in China, and the amount of information and propagation velocity are other sources can not match, the internet become one of the most important source of information. The network is also an important distribution center for information of the financial sector, especially with WEB2.0 technology development, forums, blogs, chat rooms and other interactive technologies are emerging, so that investors can participate in the process of network information creation, dissemination and access. Stock forum is one of the most popular online community, a large number of investors in the stock forum exchange information and share their experiences to assist investment decisions. Accessing to information in the online forum is an important way of understanding investor psychology and behavior.

Compared to the centuries-old mature financial markets abroad, China's financial markets still in development stage.The regulatory system is imperfect, and the majority in the market are speculators whose behavior is susceptible to the psychological impact. Get psychological and behavioral information of investors in the stock forum, and study their mutual influence on the stock market has practical significance.

Behavioral finance theory believe that the stock market will be subject to psychological and behavioral impact of irrational investors. Our study on China's stock market will base on this theory. This paper presents a method of automatically removing the domain-independent reviews, which successfully removes 84% of the stock market irrelevant reviews, and retains information for more than 90 percent of the stock market. This paper compares the semantic analysis methods, machine learning methods and N-Gram methods, and support vector machine combined with the information gain method can obtain better experimental results than the other two.

We analyze factors affecting stock prices, and develop a regression model to predict the stock price. We have analyzed single stock, the communications sector

and the Shanghai Composite Index based on stock reviews. The results showed that the lag stock's closing price, the sentiment index, institutional rating and the number of lag day news can effectively explain the closing stock price changes. With one-way ANOVA analysis on communications sector, the results show that the sentiment index is an important factor to affect the rate of return and volatility. We have done correlation analysis between sentiment index and the leading and the lagged of the Shanghai Composite Index, also done on the individual stocks. The result show that investors sentiment will be affected by the Shanghai Composite Index, and have effect on single stock price.

Keywords： Stock market; stock analysts; sentiment analysis; stock price; forecast

目　录

第1章　绪论

1.1 课题的研究背景

股票市场是国民经济状况的晴雨表，是经济发展的产物，同时又对经济发展有着重要推动作用。截止至 2011 年末，我国股票总市值达 21.38 万亿人民币[1]，占 GDP 总值的比例已达到 49.62%。成为仅次于美国的第二大股票市场，在中国和世界范围内都发挥着重要作用。

自 1811 年美国率先建立股票市场，已有二百年的历史，随时间推移，国外的股票市场已相对成熟。然而，由于我国的股票市场成立时间较短，相对于国外成熟的证券市场而言，还有许多方面不够完善，制度上的缺陷及实证工作的滞后都造成了我国股票市场的投机特点。仅有 20 多年历史的中国国内股票市场，虽然有国外的发展经验，但是由于国家政策、经济制度、国情等诸多情况不同，我国股市的发展是一个不断摸索的过程。经历了 2007 年的股市暴涨，和 2008 年底股市狂跌，在 2009、2010 年的平稳之后，2011 年从 26.29 万亿跌至 21.38 万亿元，市值跌幅达 18.67％。如此剧烈震荡的股票市场引起了投资者及股市研究者的广泛关注。

迅速发展的股票市场在国民经济中的地位日益提升，股票市场的预测、分析对经济发展具有重要意义。随着参与股市投资的股民数量迅速增长，对股票市场的详尽分析和预测的需求也日渐增加。基于以上原因，股票市场的分析和预测成为学术界研究的热点问题。在人们长期的股票研究中，基本面分析和技术分析、预测两套方法逐渐完善起来，对股票市场的运行给予了解释。在市场经济不断发展的过程中，金融市场活动日益复杂，过度反应与反应不足、羊群效应等异象不断产生。有效市场理论已不能对异象的产生给出合理的解释，在不断寻求答案的过程中，行为金融理论的提出，从投资者心理及行为的角度给出了合理的解释。非理性投资者的存在、其行为的不可预见性及有限套利性会使股票价格偏离其实际价值[2]，金融市场会受到投资者情绪的影响[3]，都是行为金融理论的重要内容。

随着互联网技术的高速发展，互联网在全世界范围内迅速普及，尤其是在中国国内，截止至 2012 年 1 月，中国网民数已达 5.13 亿，仅 2011 年全年，新增网民数就达 5500 万。互联网已成为一个各种信息的集散地。随着 Web2.0 的

流行，各种网络服务被网络用户熟知、接受、应用，人们不仅可以在网络上获取信息，而且可以随意发表信息，让众多的网络用户看到自己发布的信息，博客、论坛、微博等网络社区成为获取信息、发布信息、沟通交流的重要平台。信息是决策基础，越来越多的人以来互联网上海量信息对事物进行判断。在金融领域，网络也成为日益重要的角色，互联网不仅改变了投资者参与投资活动的方式，也成为了股票市场信息传播的主要渠道，众多的投资者通过互联网查找投资相关信息，并在网络社区与其他投资者交流投资经验。在众多额网络社区中，网络论坛是一个集中了新闻、内幕消息、个人见解的重要网络平台，论坛参与者可以发表帖子表达自己对股票市场的走势的预测、期望和见解，也可以浏览他人发布的各种信息，综合其他渠道获得的信息作出投资决策。

在众多的投资者心理、行为信息来源里，由于股票论坛信息量大、参与人数多、言论自由而受到众多研究者的青睐，早在上世纪 90 年代末，在美国等互联网及金融市场都相对发达的国家已有研究者利用股票论坛对股票市场进行研究。在已有的研究中股票评论变量主要包括股评数量及情感倾向，股票市场表现变量主要包括价格、收益、交易量、波动率等，通过对股票评论对股市表现进行分析。众多的研究证明了股票评论并非毫无意义，而是包含着能够影响股票市场的变量，因此，从股票评论中发掘股票市场相关信息，帮助投资者优化投资决策，具有现实意义，这也是本文的主要研究内容。

1.2 研究目的与意义

股票市场对推动国民经济迅速增长和世界经济一体化影响巨大，股票市场又为公司筹集资金，优化资源配置，促进公司转换经营机制的功能，从宏观方面讲，为保证国民经济健康、稳健发展，对股票市场进行监控分析、预测都是十分必要的。从微观方面讲，高回报的特征，使广大股民蜂拥而至，而高风险又让众多投资者如履薄冰，有效的对股票表现进行分析、预测成为股民降低投资风险，获得收益的必要手段，也成为了研究者的研究热点。

互联网信息被广泛应用于各个领域，本文对网络用户在股票论坛中对股票的评论进行研究，考察股票评论包含的投资者心理及行为信息，并基于该信息对股票市场进行预测。

行为金融学认为非理性投资者的心理及行为能够对股票市场产生影响，使资本价格偏离其价值，股票评论中包含了大量的投资者心理及行为信息，对其

研究具有重要意义。与国外的数百年历史的证券市场相比，我国股票市场尚处于新兴阶段，短线交易的投机者居多，会通过搜寻各类消息辅助决策。其中，股票论坛就是一种新兴的包含其他投资者提供的信息的重要平台，因此，对股票论坛中的评论进行分析具有现实意义。

以往对股票市场的研究都集中在基本面分析及技术分析上，由于股票市场异象不断涌现，从上述两方面以很难给出合理解释。从投资者的心理及行为的角度能够给出合理的解释，因而，利用股票论坛研究股票市场能够辅助投资者决策。政府部门可以通过网络论坛发掘用户的信息，辅助政策制定。在国内股票论坛的研究还非常少，本文对股票论坛中的信息探索具有实际意义。

1.3 国内外研究现状及分析

1.3.1 股票市场预测研究

自有金融市场以来，人们便开始寻求预测股票表现的方法。经过一个世纪的投资实践，已形成了基本面分析和技术分析两类预测、投资方法。

Graham 和 Dedd（1934 年）在其著作《证券分析》当中正式提出了基本面分析的概念[4]，JohnWiiliams（1938 年）提出了著名的股利贴现模型(DiscountDividendModel)，论证了证券内在价值的估计方法[5]。基本面分析主要的目标是对决定股票价格及价值的基本经济要素进行分析，包括宏观经济层面的国际贸易关系、货币政策、经济周期、通货膨胀率、失业率等，行业层面的行业生命周期、企业之间竞争、技术变革、政策影响等，以及公司层面的公司基本素质、公司财务等，来估计资产的内在投资价值，是判断中长期市场价格总体发展方向的一种分析方法。基本面分析法从股票的内在价值入手，注重公司的发展前景，是以长期投资为目标的投资分析工具。在非有效地市场中，基本分析法通常能获得异常收益。

技术分析是以凯恩斯提出的空中楼阁理论为基础，股票市场价格完全是由投资者构造出来的，完全抛开了股票的内在价值，强调心理因素对股价的决定性作用。Robert Rhea 在 1932 年《道氏理论》中提出了技术分析的完整理论体系，道氏理论指出，基础性的经济变量对证券价格的影响是一个长期过程，而证券市场作为收集所有投资者关于经济变量预期(并以此进行买卖操作)的场所，其市场指数(如道琼斯工业指数，DowJonesIndustrialAverage，DJIA)的涨跌通常领先于实际经济现象的发生[6]。此后，道氏理论经由众多研究者的拓展和

推广，最终形成了如今被广为应用的技术分析。一般来说，技术分析法可分为K线、切线、形态、波浪、指标等五类。

技术分析和基本面分析都各有优缺点，行为金融学的出现，又带来了新的投资决策方法。行为金融理论基于遵从人们的实际决策并不能遵从最优决策模型，将心理学融入到金融学之中，从微观个体的行为、心理和社会动因来了解和研究证券市场中的问题。Kahneman (1979)[7]发表的《前景理论：风险状态下的决策分析》，为行为金融学奠定了坚实的理论基础。由 Dreman 在 1981 年提出的逆向投资策略（contrarian investment strategy）[8]，利用了市场上存在反转和赢输者效应，通过买进过去表现差、卖出过去表现好的股票进行套利。Grinblatt(1995)提出惯性投资策略（momentum investment strategy）[9]，利用股票在一定时期内的价格粘性，即动量效应，通过预测价格的持续走势进行投资操作。针对投资者的损失厌恶心理，成本平均投资策略（dollar cost averaging strategy）[10]建议投资者按照预定的计划以不同的价格分批买进股票，以备不测时摊抵成本，从而避免一次性投入而造成较大的风险。针对后悔厌恶心理，及人们对风险承受能力可能会随着年龄的增长而降低的特点，时间分散投资策略（time diversification strategy）[10]，建议年轻的投资者增加股票在投资组合中比例，将债券投资比例降低，当年龄增大时，减少对股票的投资，而提高债券投资比例。芝加哥大学的 Banz 通过实证研究发现，小公司比大公司有更高的回报表现，小盘股收益率在长期内优于股票市场的平均水平，即小盘股投资策略（small company investment stratey）[11]。集中投资策略（centralization investment strategy）[12]则主张将大部分资金集中在少数几种可以在长期投资过程中产生高于平均收益的股票，坚持持股，无论股市在短期内的涨跌，直至市场发现这些股票的内在价值，股票价格得以提高。

随着不断地发展和完善，行为金融决策理论已成为成为众多投资者的分析工具，其中包括个人、机构投资者及一些证券分析师。

1.3.2 投资者情绪研究

行为金融理论有效地解释了传统金融理论无法解释的股票市场异象,为金融学的发展提供了新的理论和实践方向，投资者情绪和有限套利是行为金融学的两个最重要假设。投资者情绪是对未来现金流量的和投资风险的信念,而不是根据眼前事实进行的判断[13]。投资者的行为易受情绪的影响，通过投资者情绪

研究投资者行为，进而对股票市场进行实证研究具有理论和实践意义。

投资者情绪（Investor Sentiment）是指投资者对股票市场的预期与标准状态进行比较，其中，看涨是指投资者对收益的期望要高于平均水平，看跌是指投资者期望收益低于平均水平的收益[14]。行为金融将投资者分为噪声交易和理性套利两类,并且股票市场会受到投资者情绪的影响[3]。投资者情绪的度量主要有两种方式，一类是直接搜集法，向投资者群体发放问卷的调查结果编制而成，此类指标有股市信心指数（Stock MarketConfidence Indexes）、卖方指标、投资者智慧指数[15]、美国个人投资者协会指数(AmericanAssociation of Individual Investors)[16]等；另一类是间接情绪指标，该指标主要是通过对股票市场交易数据进行处理，提取出反应投资者情绪的指标。此类指标有腾落指数(ADL)及其变形、IPO 当日收益率、封闭式基金折价率[17]、道富投资者信心指数(State Street Investor Confidence Index)[18]、波动率指数 VIX(Volatility Index)[19]、认沽认购比例 PCR(Put/Call Ratio)[20]、以及股票市场投资者情绪 EMIS(Equity Market Investor Sentiment)[21]等。

大量研究从投资者的角度，应用反应投资者心理的指标分析投资者情绪与股票市场的关系，并预测股票市场表现。Neal 分析了美国证券市场 1933 至 1993 年之间的数据，证明了开放式基金净赎回、零股买卖比、封闭式基金折价三个投资者情绪指标可以预测规模溢酬及企业间的收益差异[22]。Baker（2006）利用主成分分析方法，从多个情感指数中提取了主要情感指数，通过分析表明其与股票横截面收益相关[23]。杨春鹏等（2009）将央视看盘数据作为机构投资者情绪指标，研究发现投资情绪对沪深两市收益率有显著影响，而两市的收益率也是影响投资者情绪的重要因素[24]。池丽旭等将投资者情绪分为理性情绪和非理性情绪，发现在股权改革之前，不同投资者情绪对股票收益产生不同的影响，改革滞后，影响不够显著[25]。众多研究都从投资者的心理入手，对股票市场进行研究，尽管研究结果不同，但是为解释股票市场变化提供了新的方法。

1.3.3 股票评论研究

在线股评是互联网用户对在网络上对股票的进行评论，包括股票论坛、股票留言板、股票博客、股票微博等网络平台上由网民发表的关于股票价格、趋势、内幕消息、新闻、政策等各方面信息的评论。早在 20 世纪 90 年代网络平台就被用来进行股市信息交流，并且一些研究者已经开始利用网络评论对股票

市场进行研究。部分研究者机遇股票论坛内股评数量及投资者情绪对股票市场进行分析，大部分研究证明股票评论对股票市场变量有解释作用。

Wysocki [26.27]利用 Yahoo！Finance 上的股票论坛，通过截面分析和事件序列分析的方法证明了在线股评数量可以预测次日的股票市场交易量和股票收益。而 Tumarkin [28]研究了 RagingBull 上的股评数据，发现股评活动与股票收益率和交易量不存在关联关系。Antweiler 在 2002 年同样通过 Yahoo！Finance 股评数据的研究发现，股评数量和股市波动、股市收益存在相关关系[29]，在 2004 年分别利用了 Yahoo！Finance 和 RagingBull 上的数据证明了在线股评数量可预测股票交易量和波动。Sabherwal [30]在 2008 年通过事件研究和回归分析的方法分析了 TheLion 上的股评信息，发现在线股评数量与异常收益有关，在股票成为讨论热门股的当天，股票会产生 19.4%的异常收益。

Das 在 2005 年开始研究在线股评所表达的情感信息，通过相关分析和回归分析得出结论，在线股评所表达的情感与股票收益相关，但不能对股票收益进行预测[31]。在 2007 年发现前一天在线股评所反应的情感综合指数与技术部门总指数显著正相关，但对单只股票这个结论并不成立[32]。Sehgal 在处理论坛数据时映入了评论作者的可信度作为情感指数影响作用的权重，发现情感指数与股票表现情况相关[33]。Zimbra 利用沃尔玛公司论坛数据，结合主题提取和文本分类的方法，通过回归分析发现论坛内股评反映的情感与话题可以预测股票价格[34]。Koski [35]发现股票评论与股票市场波动存在因果关系，股票评论能够引起股市波动。通过对 Yahoo！Finance 论坛数据及 Stocktwits 微博数据的分析，Chong 证明了微博观点对股票市场的运动趋势存在预测作用，并且证明了不充分反应现象的存在[36]。

1.4 本文的主要研究内容

本文的研究内容主要有：应用为本分类及情感分析技术提取股票论坛中股票市场相关数据，并挖掘在股票评论中投资者观点（看涨、看跌）；应用股票论坛数据(评论数量分布、评论观点)对股票表现变量（换手率、收益率、交易量、波动率）相互影响进行研究；建立股票价格预测模型，对股票价格进行预测。研究框架如图 1-1 所示。

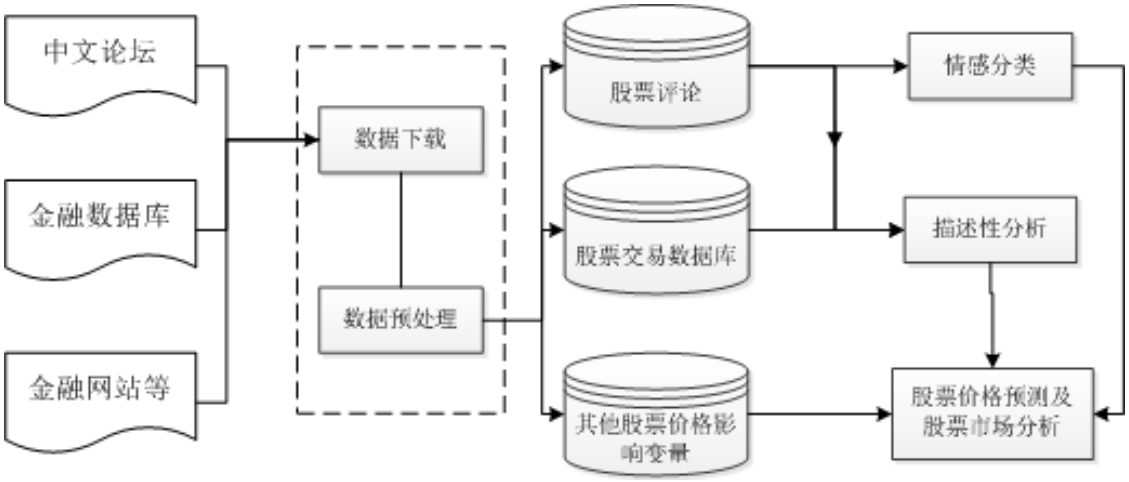
中文论坛
金融数据库
金融网站等
数据下载
数据预处理
股票评论
股票交易数据库
其他股票价格影响变量
情感分类
描述性分析
股票价格预测及股票市场分析

图 1-1 研究框架

第2章　理论与方法基础

2.1 行为金融理论概要

有效市场假说认为投资者完全理性，投资行为是随机的，而行为金融学对此提出了质疑。行为金融学认为投资者并非理性，并认为投资者的心理、情绪会对其行为、决策产生影响，进而影响股票表现及金融市场。行为金融学主要的研究内容是，投资者在参与金融活动过程中的认知、情感及行为特征，以及其对股票市场的影响。行为金融学大师 Fuller 行为金融学至少具备以下特点：是一个综合学科，融合了社会学、心理学以及决策科学等标准理论；试图解释有效市场理论中无法有效解释的金融市场异常；研究投资者判断系统性错误的方式[37]。

经过多年来的发展，行为金融理论逐渐确立了自己立论的最重要的理论基石，有限套利和投资者心理。由于金融产品的不完全替代，套利者分先厌恶以及噪声交易对股票价格产生干扰，使得套利者无法实现充分套利。人们在决策的过程中存在心理偏差和偏好，而这些偏差和偏好会对现实世界中的证券买卖产生影响，也就是投资者心理，行为金融学的另一个重要理论。

行为金融学研究的核心内容概括为以下几个方面：金融市场中的异象，心理偏差与非标准偏好以及有限套利[38]。

2.1.1　股票市场异象

1．过度反应和反应不足

过度反应是由于投资者倾向于重视新的消息而忽略旧的消息，使得投资者对最近的股票价格变化赋予过多的权重。因此，股票价格在利空消息下下跌的过度而在利好消息下上升的过度[39]。

反应不足是指当影响股票价格的消息出现之后，最初股票价格未能充分体现该消息对其产生的影响，经过时间的调整，消息的影响才逐渐完全体现在股票价格上[39]。

保守主义和代表性启发是造成这两种现象的一个重要心理因素[40]。代表性启发法使投资者对近期数据变化赋予过多的权重，而忽略数据整体特征，也会

使投资者仅依赖一部分信息作出决策，最终导致投资者对该部分信息的反应过度。保守主义则使投资者过分依赖过去的信息，不能及时根据最新信息调整判断和决策，而导致反应不足。多数情况下，投资者会对相对容易处理的信息过度反应，对不易获取的数据反应不足。

过度自信和自我归因偏差是导致过度反应和反应不足的另一重要的心理和行为因素[41]。过度自信会使投资者夸大自己对股票价值判断的确定性；自我归因偏差则导致投资者低估关于股票价值公开信息对股价的影响。随时间推移，自我的行为偏差会被公共信息所战胜，对信息的过度反应和反应不足，最终导致股票回报在短期具有连续性，在长期内会有反转作用。

2．噪声交易

噪声交易是行为金融学的主要内容之一。“噪声”在金融领域是指与投资价值无关的信息[42]。这种信息可能是股市参与者故意制造以迷惑其他投资者的信息，也可能是因为判断错误而产生的信息。噪声交易者是指基于噪声进行交易的投资者。噪声交易者在不发达的以及发达的金融市场普遍存在，噪声交易会导致股票投资系统风险增大，进而导致总风险增大[43]。

行为金融理论将投资者分为理性交易者和噪声交易者。理性交易者基于股票的基本价值进行交易，噪声交易者基于噪音进行交易。噪声交易者使股票价格偏离基本价值，而理性交易者能够消除这种影响，使股价回到基本价值。投资者情绪高涨或低落时，噪声交易都会被放大，使得股票市场的流动性增大，从而给股市带来强烈的波动。投资者情绪会对股票价格产生系统性影响，与股票的超额收益呈正相关，其变化对导致股票收益的变化[21]。

2.1.2 心理偏差与非标准偏好

经典金融学中理性投资者满足两个假设：（1）投资者可以对未来股票未来的状态做出无偏估计，并按照贝叶斯法则处理新的信息；（2）投资者对风险的偏好符合预期效用理论。

在行为金融学中，人们在形成预期的过程就是判断某件事情发生的概率，在经济学或心理学实验证明了心理偏差会对金融决策行为产生影响，例如，可能性偏差、代表性偏差、锚定、过度自信等。行为金融学运用这些偏差来解释金融市场异象，提出了 BHS、BSV、HS 模型等。

由于投资者根据自己的预期进行投资决策，能够掌握投资者心理，并发掘

投资者的预期模式是投资者对研究金融市场变化有重大意义。

2.2 情感分类

由于主观性文本中含有大量的有价值信息,文本的情感分析也受到了众多研究者的关注。情感分析是分析给定文本片段中所包含的说话者的情感倾向，分析结果通常分为正面，负面和中性三类[44]。在本文中所处理的文本是在线股票评论，以下为三类情感倾向代表句子：

正面---“预测会沿着 5 天均线上行，反弹预计会创 3123 点以上的新高”

负面---“快逃命啊！明天就会百多点的大跌、暴跌----!!!! 快逃命啊!”

中性---“权重股维稳迹象明显，由于总体成交量表现不理想，建议继续观望”。

目前情感分析技术主要包括语义分析方法（Semantic Orientation）[45]及机器学习方法（Machine learning）[44]。

2.2.1 应用语义分析进行情感分类

语义分析法是通过分析文档中每个词的情感倾向,为获得整篇文档中的情感倾向。Turney[45]在 2002 年提出了基于 PMI-IR 算法的语义情感分类的方法，他将点互信息法（Point Mutual Information，PMI）与信息检索法（Information Retrieval，IR）结合起来，利用网络搜索引擎后天数据库计算每个词语的语义倾向信息，进而对词语的情感倾向进行判断。

Turney 提出的语义情感分析法（SO-A，Sentiment orientation from Association）主要是通过分析词语间的关联程度来获得词语的语义倾向。其中，逐点互信息分析（SO-PMI，Semantic Orientation from Pointwise Mutual Information）是一个重要的方法，通过统计词语间的共同出现的次数来计算词语间的关联程度，定义如下。

$$\mathrm{PMI}\left(term_1,term_2\right)=\log_2\frac{\left(\Pr\left(term_1\cap term_2\right)\right)}{\Pr\left(term_1\right)\Pr\left(term_2\right)} \tag{2-1}$$

其中，$\Pr\left(term_1\cap term_2\right)$是$term_1$与$term_2$同时出现的概率；

$\Pr\left(term_1\right)$，$\Pr\left(term_2\right)$分别是$term_1$和$term_2$出现的概率。

利用SO-PMI方法，英文中的一个词汇的语义倾向可以如下下式进行计算：

$$\mathrm{SO}(term)=\mathrm{PMI}(term,"excellent")\text{-}\mathrm{PMI}(term,"poor") \tag{2-2}$$

式中的“excellent”，“poor”是作为语义倾向的参考词汇的，英文中的5星评论系统中，将一颗星评为“poor”，五颗星评为“excellent”，当term语义更靠近“excellent”时，$\mathrm{SO}(term)$为正，而当term语义更靠近“poor”时，$\mathrm{SO}(term)$则为负。在一些研究中为了估计词语之间共现的概率，采用信息检索（IR，Information Retrieval）的方法，通过搜索引擎搜索的结果数来估计词语共现的概率[45]。

基于以上方法，结合中文文本的特点，可通过如下步骤对中文文本进行情感分类：

步骤一，对中文文本进行词语划分，并对每个词语进行词性标注；

步骤二，分析中文表达模式，选出具有语义倾向的词语模式；

步骤三，选出文档中可以表示极端正面和极端负面的代表词；

步骤四，利用一对代表极端正面和极端负面情感的标准参考词(Reference Words Pair, RWP)，结合baidu等网络搜索引擎，计算步骤二中每个具有语义倾向的词组的语义倾向值，即SO值；

$$SO(phrase)=\log_2\left(\frac{hits(phrase\cap rw-p)hits(rw-n)}{hits(phrase\cap rw-n)hits(rw-p)}\right) \tag{2-3}$$

在这里$hits(x)$表示字符串x在搜索引擎中查询，所返回结果的总页数，例如，$hits(rw-n)$表示利用该搜索引擎查询负面代表词时返回的结果页数，$hits(phrase\cap rw-p)$表示该词与正面参考词再搜索引擎查询时同时返回的结果页数。如果一个词组与正面代表词共同出现的概率大于与负面参考词共同出现的概率，则该词组为正面情感词组的可能性大于负面参考词的可能性。

步骤五，计算一个待分析的文档中所有具有语义倾向的词组的的SO平均值，并以特定阈值为判断标准，最终确定文档的情感倾向。

2.2.2 应用机器学习进行情感分类

机器学习是研究利用计算机识别现有的知识，并获取新知识和技能的学科，就是利用计算机模拟人类学习活动的过程。情感分类的另一种方法就是将其看做基于主题的文本分类问题，即机器学习的一个分支。在文本分类中表现良好的 naive Bayesian、SVM、kNN、决策树等算法也可应用于情感分类。该方法主要依靠词语间相似度或文档中词频数进行分析，通过训练样本文档特征进行模型训练，计算出相关类别各词语出现特征，再结合目标文档特征判别其类别。该方法已应用在电影评论、餐厅评论、宾馆评论、股票评论中，并取得了良好的分类效果。鉴于中文不同于英文的特点，中文机器学习方法首先需要进行中文分词，具体的中文情感分类步骤如下：

(1)文本分词

(2)文本的特征表示

(3)特征选择

(4)分类模型

(5)分类模型评价

2.2.2.1 文本分词

随着计算机技术，尤其是网络技术的迅速发展和普及，人们迫切希望计算机能够像人类那样对自然语言进行处理。词是最小的能够被独立运用的语言单位，而中文文本不同于西文文本中有空格显示地将词划分开来，使得计算机处理中文要难于对西文的处理。因此，中文的自动分词就成了计算机处理中文文本时的一项基础性工作。

汉语自动分词的困难主要有三方面：首先是汉语分词规范问题，由于单个词与词素之间的划界以及词与短语之间的划界不清晰而导致汉语分词并不规范[46]；第二是歧义切分问题，由于在汉语中普遍存在歧义字段，歧义切分成为中文自动分词重大阻碍；第三是未登录词问题，这部分包括不断涌现出的普通词汇或专业词汇以及专有名词（包括人名、地名、外文译名、组织机构名等）。

根据各分词方法原理的不同，可以将其分为三类，分别是基于词表匹配的分词方法、基于理解的分词方法及基于统计。以下将分别对各方法原理进行简单介绍。

基于词表匹配的分词方法，又叫机械分词方法，该类方法需要有一个庞大的词典做支撑，按照一定的策略将待分析文本与该词典进行匹配，若匹配成功，

则找到一个词。根据匹配策略的不同，有以下几种方法，分别是正向及逆向最大匹配法、最少切分、双向最大匹配法等。

基于理解的分词方法，该方法的基本思想是对待分析文本进行句法、语义分析模拟人对句子的理解来处理歧义现象。该方法主要包括总控部分、句法语法分析和分词三部分。由于该方法需要大量的语言知识和信息，还处于试验阶段。

统计分词方法，是基于相邻的字同时出现的概率越大，就越有可能构成一个词的原理而实现的。通过统计语料中相邻字共现的频率，计算互信息值，通过互信息值来反应词之间的紧密程度，进而判断是否属于同一个词。典型分词方法有，基于 HMM（隐马尔可夫模型）分词方法，基于三元统计模型的分词方法等。

2.2.2.2 文本的特征表示

为了计算机能够识别文本，在进行中文处理之前必须将为本表示成计算机可以是别的形式。向量空间模型(Vector Space Model,VSM)，又称文本的词袋(BOW)表示是目前最常用的文本表示方法，是 Salton 等人于上世纪 60 年代末提出的，并在 SMART(System for the Manipulation and Retrieval of Text)信息检索系统中得到了成功应用[34]。

在该模型中，一个文档被看做 n 维向量空间中的一个向量，特征项可以是字、词、词组或短语等，是 VSM 中最小且不可分的语言单元，每个特征项被看做向量空间的一个维度，在一篇文档里，为表示每个特征项在文档中的重要程度，都被赋予一个权重。根据权重的不同，可以分为布尔表示法、词频表示法等。

除了 VSM 以外，还有词组表示法、概念表示法等文本表示方法。虽然词组表示法提高了特征向量的语义含义，但却降低了他的统计质量，导致这一方法在文本分类中未得到显著地效果[47]。概念表示法中用概念作为特征向量的特征表示法，但需要额外的语义词典资源，有益于文本分类效果的提高[48]。

2.2.2.3 特征选择

在向量空间模型建立之后，所有词都作为特征来表示文档，但并非所有的特征都是有效的，即能够提高分类准确率。过多的特征会导致分类模型训练速度降低，分类的准确率也会因为噪音特征的影响而降低。目前为止，已有较多的特征选择算法，基本思想都是采取一定的策略评估各个特征对分类的重要程

度，依照重要程度的不同，对各个特征进行取舍从而达到特征降维的目的。最常见的特征算则方法如下所示：

（1）文档频率法 文档频率（DF）是指该特征曾在多少个训练文档中出现过。首先统计出训练语料中某个特征的 DF 值，根据设定的阈值，去掉 DF 值小于该阈值的特征，保留剩余的特征作为特征选择结果。

（2）信息增益（IG）法[49] 通过计算某个特征在分类过程所能够产生的信息量，设定一定的阈值，信息量大于该阈值的特征项保留，否则删除此特征项。信息增益的计算方法如式 3-4 所示。

$$G(t_i)=\left\{-\sum_{j=1}^{M}P(C_j)*\log P(C_j)\right\}- \left\{\begin{array}{l} P(t_i)*\left[-\sum_{j=1}^{M}P(C_j\mid t_i)*\log P(C_j\mid t_i)\right]+ \\ P(\overline{t_i})*\left[-\sum_{j=1}^{M}P(C_j\mid \overline{t_i})*\log P(C_j\mid \overline{t_i})\right] \end{array}\right\} \tag{2-4}$$

其中，$P(C_j)$为C_j类文档在训练样本中出现的概率，$P(t_i)$为训练文档中包含特征t_i的概率，$P(C_j\mid t_i)$为文档中包含t_i时属于类别C_j条件概率，$P(\overline{t_i})$为样本中不包含t_i 的概率，$P(C_j\mid \overline{t_i})$是文档不包含特征项$t_i$时属于$C_j$的条件概率，M 表示类别数。

（3）χ^2统计量（CHI） 通过度量类C_j与特征项t_i之间的关联程度，而决定该特征项的取舍，假设二者符合一阶自由度的χ^2分布[50]。若某特征对类C_j的 CHI 值越大，表示它对该类别的识别作用越大，反之则越小。

$$\chi^2_{MAX}(t_i)=\max_{j=1}^{M}\left\{\chi^2(t_i,C_j)\right\} \tag{2-5}$$

（4）互信息法（MI） 互信息越大，特征t_i与类别C_j共同出现的概率越大，二者的互信息可由下式计算：

$$I(t_i,C_j)=\log\frac{P(t_i,C_j)}{P(t_i)*P(C_j)} \tag{2-6}$$

为了选出对多类文档识别有用的特征，特征t_i的类别判断如下：

$$I_{MAX}(t_i)=\max_{j=1}^{M}\left\{P(C_j)*I(t_i,C_j)\right\} \tag{2-7}$$

2.2.2.4 文本分类

文本分类是一个有监督的学习过程，主要包括训练和预测两个部分。在训练过程中，分类器读入训练数据，并对训练文本的分类模式进行学习，建立分类模型。在预测过程中，将待分类的数据输入分类模型，分类模型根据已学习到的模式对待分类文本进行类别预测。经过长期的研究，已有多种分类模型被发明及应用：主要有相似度模型（KNN）、概率模型（贝叶斯）、线性模型（支持向量机）及非线性模型（神经网络、决策树）等。

(1) 支持向量机

支持向量机(support vector machine，SVM)是基于统计 VC 学习理论及结构风险最小化（SRM）原理的机器学习方法。具有理论严密、适应性强、全局优化、训练效率高和泛华性能好等优点，能够非常成功地处理模式识别（分类、判别分析）和回归问题（时间序列分析）等诸多问题，并可进一步推广到预测和综合评判等领域。SVM 能够根据有限训练样本建立的有效的分类模型，并且对独立的测试样本集仍可保证较小的误差[51]。SVM 作为一个非常优秀的分类器被广泛应用于文本分类领域。由于本文主要应用 SVM 进行文本分类，所以一下部分主要对支持向量分类机进行介绍。

SRM 是针对二值分类问题提出的，因此，SVM 的基本问题也是二值分类问题。设定线性样本集$(x_1,y_1),(x_2,y_2),...,(x_m,y_m),x\in R^n,y\in\{+1,-1\}$是类别标号，$X$为具有 n 个属性的向量。要求在$X=R^n$上找到$g(x)$，得到决策函数，即分类学习机器

$$f(x)=\mathrm{sgn}(g(x)) \tag{2-8}$$

存在一个超平面如公式 2-9 所示，其中$x\in R^n,w,b\in R^n$使得样本集中任意的一组数据(x_i,y_i)，满足公式 2-10。

$$w^T \cdot x + b = 0 \tag{2-9}$$

$$y_i \left[w^T \cdot x + b \right] - 1 \geq 0 \quad i = 1, 2, ..., m \tag{2-10}$$

支持向量机的原理可以表示成如图 2-1 所示。

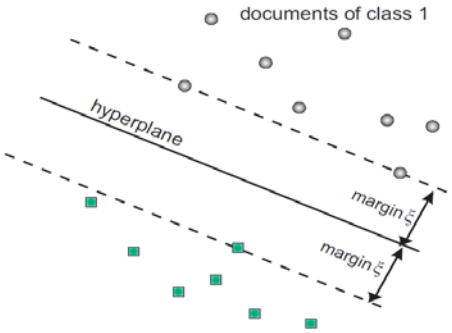

图 2-1 支持向量机原理

由图 2-1 可以看出，利用 SVM 分类，就是寻找一个超平面，使得两个样本中离分类面最近的点与分类面的距离最大，即，使得分类超平面两侧的空白区域（margin）最大，这个平面就是最有分类超平面。在两个样本中离最优分类超平面最近的点且平行于最有分类超平面的 $X_{(1)}$ 和 $X_{(2)}$ 上的训练样本（2-10）中等号成立的样本，这些样本就是支持向量。当要确定一个待测样本是属于哪一类时，判断 $w^T \cdot x + b$ 的值大于 1 属于类别 1，否则属于类别-1。

（2）朴素贝叶斯分类器

朴素贝叶斯(Naive bayes)根据特征项和类别之间的关联概率来估计文档所属类别的概率。假设文本是基于词的一元模型，词之间是独立的，其出现与否不依赖于其他词及文档的长度，而是依赖于文档的类别。根据贝叶斯公式，文档 Doc 属于类别 C_j 的概率是：

$$P\left(C_j \mid Doc\right) = \frac{P\left(Doc \mid C_j\right) * P\left(C_j\right)}{P\left(Doc\right)} \tag{2-11}$$

在已有研究中证明，贝叶斯分类器的性能易受分类任务的影响，分类结果并不理想[52]。

（3）K-最邻近算法（KNN）

KNN 的基本思想是对于一个特定的待分类文档，分类系统会计算训练集中

每个文档与该文档的距离，并选择其中距离最小的 k 个文档对该待判别文档的所属类别进行判定。把待分类文档与临近文档的相似度作为权重，将属于某一类别的临近文档的权重求和，作为待分类文档与该类别的相似度。将该文档与个候选类别相似度排序，将该文档划入相似度最高的候选类别。相似度计算入式 3-12 所示。

$$y\left(\vec{x},C_j\right)=\sum_{d_i\in KNN} sim\left(\vec{x},\vec{d_i}\right)y\left(\vec{d_i},C_j\right)-b_j \tag{2-12}$$

其中，$y\left(\vec{d_i},C_j\right)$取值为 0 或 1，分别表示文档$\vec{d_i}$属于和不属于类别$C_j$的情况，$sim\left(\vec{x},\vec{d_i}\right)$表示待测试文档$\vec{x}$和训练文档$\vec{d_i}$之间的相似度，$b_j$为阈值。

文本分类技术发展至今，除了上述介绍的几种分类算法外，还有神经网络法（neural vector machines，NNet）、决策树法（decision tree）、模糊分类法（fuzzy classifier）等算法。

（4）基于 N-gram 模型的分类算法

N-gram 模型每个词出现仅与其前面的 N-1 个词有关，通过最大似然法估计前 N-1 个词出现对第 N 个词出现的概率，成功避免了中文文本处理中文分词步骤。基于 N-gram 的分类算法是利用 N-gram 模型提取特征词，通过训练样本进行学习得到分类模型，在利用该模型对待分类数据进行类别预测。

2.2.3 分类模型评价

对于一个分类器的评价指标主要有两个，首先是分类准确率，其次是分类的效率，包括时间效率和资源耗费。本文主要应用第一个指标分类模型进行评估。在建模的过程中需要对训练数据集进行学习，在模型评估的过程中需要将建立好的模型应用在测试数据集上，将分类器预测的结果与正确结果对比分，评估模型，因此，这两个数据集都需要具有正确的类别标定。评估模型的标准有很多，不同的标准会对模型不同方面信息有所侧重。对于一个分类器的分类结果可表示如表 2-1 的矩阵形式。

表 2-1 文本分类器输出结果

	属于类别X	不属于类别X
分类器标记为类别X	A	B
分类器标记为非类别X	C	D

其中，A表示将本身属于类别X，且被分类器分至类别X的文档个数，B表示本来不属于类别X但却被标记为类别X的文档数量，C表示属于X类别，却被标记为其他类别的文档的个数，D表示本来就不属于类别X，在分类过程也没有被分到类别X中的文档数量。该分类器对于类别X有以下评估标准，召回率（Recall）如公式3-13所示、正确率（Precision）如公式3-14所示，F-测度值（F-measure）如公式3-15所示。

$$\mathrm{Re}call_X = \frac{A}{A+C} * 100\% \tag{2-13}$$

$$\mathrm{Pr}ecision_X = \frac{A}{A+B} * 100\% \tag{2-14}$$

$$F-measure_X = \frac{2 * \mathrm{Pr}ecision * \mathrm{Re}call}{\mathrm{Pr}ecision + \mathrm{Re}call} \tag{2-15}$$

为综合评价模型对所有类别的评估结果，本文引入了微平均[53]作为评估指标，计算结果如下：

$$Mic\_Avg = \frac{A+D}{A+B+C+D} \tag{2-16}$$

2.3 本章小结

本章主要对行为金融理论进行了概括，并介绍了情感分析技术，为我国股票论坛的研究提供了理论及技术基础。股票市场异象丛生，有效市场理论已难以给出合理解释，行为金融理论的出现填补了这一空白，从投资者心理及行为的角度对其做出了解释。本文利用股票论坛中包含的投资心理及情绪信息对股票市场进行研究是有实践意义的。情感分析技术应用于股票评论的研究在国内还比较少，本文的尝试是十分有意义的。

第3章　股票评论情感分析

3.1 数据收集及预处理

3.1.1　数据来源

本文主要的研究对象是网络股票论坛内的股票评论，包括股票评论的数量及情感倾向信息。在国外的相关研究中，主要采用了 Yahoo!，Ragingbul 等大型网站的金融板块作为数据来源。在国内也不乏类似的大型网站，例如新浪，搜狐等大型门户网站的财经板块都提供了个股股吧等股票论坛，东方财富网、金融界等财经类网站也提供了关于个股及大盘的股票论坛。各网站都将股票按照名称或代码对上市公司信息进行汇总，包括公司经营状况、新闻公告、股票等相关信息便于投资者查询，并为各只股票开辟了相应的讨论板块（股吧）。

本文依据 Alexa 中国官方网站（http://cn.alexa.com/，2012 年 3 月）提供的金融类网站排名，结合各网站股吧板块的访问量等信息对综合排名前 15 位的网站进行比较，排除银行类网站及不含股票类网站主要剩余如表 3-1 所示的四个网站。

表 3-1 金融类网站比较

网站	网址	人均页面浏览量	股吧日均 IP 访问量	财经类网站综合排名
新浪财经	http://finance.sina.com.cn/	5.4	124872	1
东方财富网	http://www.eastmoney.com/	5.8	433884	2
搜狐财经	http://business.sohu.com/	2.25	346289	9
金融界网	http://www.jrj.com.cn/	3.1	30186	11

其中新浪财经和搜狐财经是门户网站，分别在财经板块提供了股票论坛，而东方财富网及金融界都为专业的金融类网站，均分别设置了股票论坛。新浪财经虽然综合排名最靠前，但大部分访问者都是为了浏览财经类新闻，而股吧的访问量不大。东方财富网综合排名仅次于新浪网，但从股吧访问量，人均页

面浏览量来比较都是最优秀的，所以将它作为首选数据来源。搜狐财经的股吧访问量较高，近搜狐财经的 3 倍，但人均页面浏览量不高。金融界综合排名最差，而且股吧访问量也为最小，不作为考虑对象。根据本文的研究需要，根据本文的研究需要，选择新浪财经、东方财富及搜狐财经作为数据来源。

本文将上海证券交易所数据作为股票市场分析数据来源，主要从国泰安数据库及搜狐股票历史数据板块获得，主要包括以下字段，股票开盘价、收盘价、最噶婆家、最低价、成交量及成交额等。

3.1.2 数据下载

文本评论作为本文的主要数据，利用 Java 程序进行数据下载，通过 SQL sever 2005 数据库进行数据存储。首先，编写 Java 程序连接网站服务器，进行网页读取，并以文本文件的形式保存到本地。由于数据量较大，采用多线程方式以充分利用计算机、网络资源，提高下载效率，对于一次下载失败的网页，进行多次重新下载以避免数据丢失，影响实验结果。再利用 java 程序解析存储到本地计算机的文本数据，移除网页中的标签等数据，提取本文实验所需评论数据，按照评论的不同分别存入 SQL sever 2005 数据库中，用于后续分析。保留的字段包括：评论主题、内容、作者、发表时间、点击数、回复数、作者名（或 IP）、作者等级等。

本文实验过程中所使用的股票市场数据、宏观经济数据主要来自于国泰安数据库，部分数据来自东方财富网的历史数据板块。

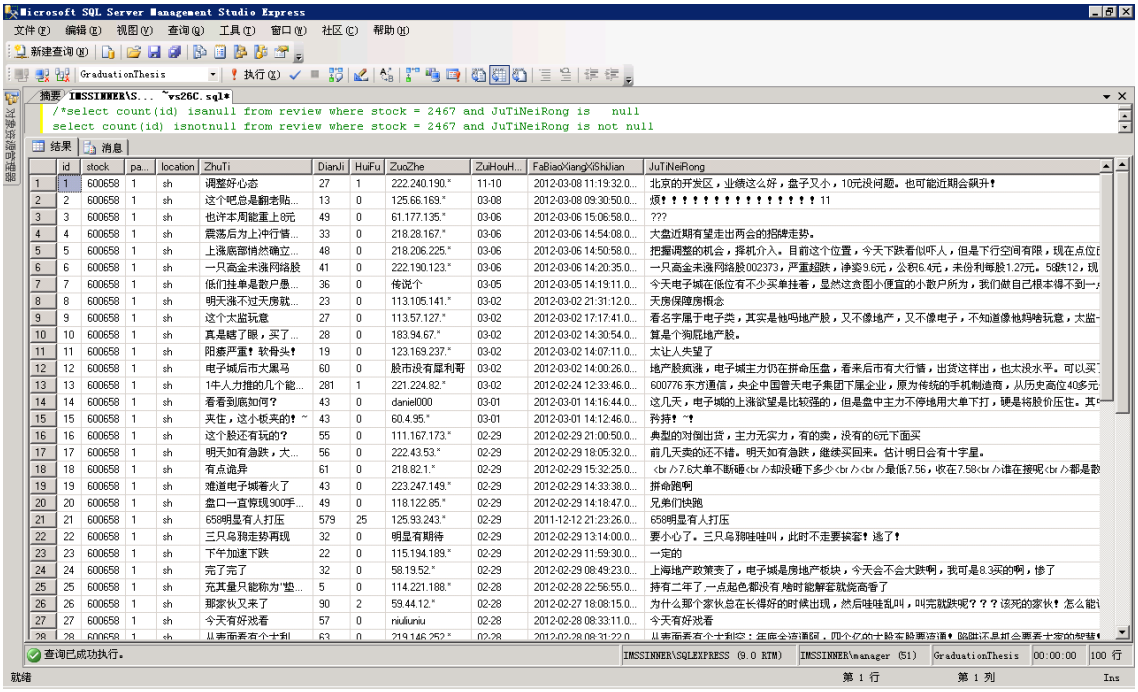

图 3-1 数据存储

3.2 无关数据清除

由于股票评论中存在大量的与股票无关信息，例如，在 1000 条训练数据中有 363 条是股票无关数据，无关数据涉及的内容广泛，在对评论数据进行分类之前无关评论进行清除是十分必要的。本文提出了基于文本分类的无关数据清除方法，基本框架如下所示：主要包括三部分：数据获取、特征生成、分类与评估三个步骤完成。

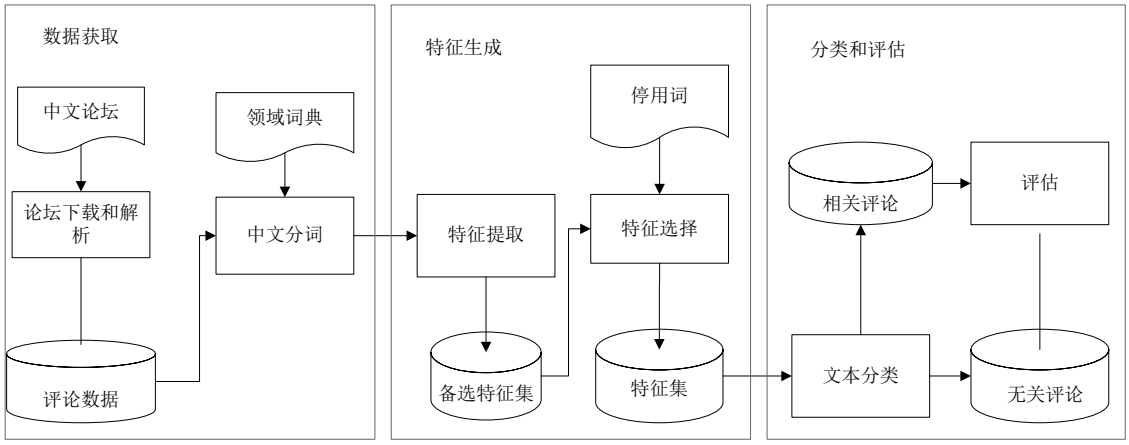

图 3-2 无关数据清除流程

3.2.1 数据获取

此步骤包括中文论坛网页下载、网页解析存储和中文分词几个步骤。网页的下载及计息存储都通过 Java 程序实现，中文分词采用中科院分词程序包完成。

3.2.1.1 论坛数据特点描述

文本部分利用了东方财富网，谈股论金板块的关于上证指数的评论作为数据来源，该板块评论未被明确标定评论者情感倾向，需要人工标注，由于数据量较大，并且需要选择有代表性的数据，从近 30000 条评论中随机选择 1000 条进行人工标定，作为实验数据。人工标定的过程中，这 1000 条数据中的 363 条被标记为股票市场不相关数据，剩余 637 条被标记为股票市场相关数据。以下是四条在论坛内的评论，其中前两条是股票是长相关评论，后两条是股票市场无关数据，即本实验要情清除的数据对象。

- 九月将是艳阳天.今天大盘走的死气沉沉, 只不过是为了做月线收盘价而已, 明天将在权重股的带领下重拾升势,开启金九升途。
- 预测会沿着 5 天均线上行，反弹预计会创 3123 点以上的新高
- 关公面前耍大刀，来呀，温酒。。。最近在读三国，新三国拍的真不好看。
- 大碗茶聊吧-----请进(2010.08.20）大碗茶聊吧把“快乐投资、快乐投机、快乐友情、快乐人生”作为办吧主题思想，为朋友们提供一个敞开心扉、互相交流、诉说甜酸苦辣的平台。

3.2.1.2 中文分词

汉语词法分析系统（Institute of Computing Technology, Chinese Lexical Analysis System，ICTCLAS）是由中科院计算机技术研究所研究开发的，经过了近十年的不断优化、更新，主要基于层叠隐马尔可夫模型（Hierarchical Hidden Markov Model），具有中文分词及词性标注、新词识别，命名实体识别等多个功能。分词正确率高达 97.58%。分词和词性标注处理速度为 543.5KB/s。ICTCLAS 全部采用 C/C++编写，支持主流的操作系统，如 Linux、FreeBSD 及 Windows 系列操作系统，支持 C/C++/C# /Java/ Delphi 等开发语言。

本文利用 ICTCLAS 系统提供的 API，通过 Java 语言进行开发，对在线股票评论信息的具体内容进行中文分词，并标注词性。为提高金融领域词汇的识别能力以及后续处理分析的准确性，本文在分词的过程中加入了一个金融词库。该词库由中国人民大学信息学院经济信息管理系梁循教授的提供，涵盖了 2 万多条金融领域的常见词汇。如下为词库中部分代表性词汇。

金融词库词汇：利多, 利空, 多头, 空头, 反弹, 盘整, 死多头, 多翻空, 短多, 斩仓, 割肉, 套牢, 多杀多, 热门股, 对敲, 筹码, 踏空, 跳水, 诱多, 骗线, 阴跌, 停板, 洗盘, 平仓, 平开, 高开, 内盘, 均价, 填权, 多头陷阱, 溢价发行, 场内交易, 散户, 建仓, 总市值, 实多, 浮多, 回档, 盘档, 抢帽子, 跳空, 补空, …。

本文将加入金融词库和未加入金融词库的分类结果做了对比，结果如下所示，

表 3-2 评论内容

	内容
评论	预测会沿着 5 天均线上行反弹预会创 3123 点以上的新高
未加入金融词库的分词结果	预测/v 会/v 沿着/p 5/b 天/n 均/ag 线/n 上行/vn ，/n 反弹/vi 预计/v 会/v 创/vg 3123 点/t 以上/f 的/ude1 新高/n
加入金融词库的分词结果	预测/DICT 会/v 沿着/p 5/b 天/n 均线/DICT 上行/DICT ，/n 反弹/DICT 预计/DICT 会/v 创/vg 3123 点/t 以上/f 的/ude1 新高/n

其中标记为 DICT 的为金融词汇库内收录的词汇。可以看到，“均线”是

金融领域常用词汇，是移动平均线指标的简称，是反映价格运行趋势的重要指标。在未加入金融词库时 ICTCLAS 未将该词识别出来，再加入该词库之后，成功识别出该词汇，而“预测”、“上行”、“反弹”等在加入金融词库前后都能被识别出来。

3.2.2 特征生成

3.2.2.1 特征表示

经过中文分词之后，统计出 1000 个文档内的所出现词汇数为 19,904 个，利用目前研究中最常用的向量空间模型（VSM）将所有文档表示成向量空间中的一个点，每个词汇都是向量空间的一个维度，每个维度的权值采用 tf-idf(term frequency–inverse document frequency)[54]计算得出。这样就得到了实验文档的可以被分类器识别的表示模型。tf-idf 权值计算方法如下：

$$tf-idf=\frac{tf_{it_k}*idf_{t_k}}{\sqrt{\sum_{k=1}^{n}\left[f_{it_k}*idf_{t_k}\right]^2}} \tag{3-1}$$

$$tf_{it_k}=\frac{freq_{it_k}}{\max_1 freq_{it_k}} t_1\in\{t_1,t_2,...,t_n\} \tag{3-2}$$

$$idf_{t_k}=\log\left(\frac{N}{n_k}+0.01\right) \tag{3-3}$$

其中，n 为所有特征的个数，n_k 为含有特征 t_k 的文档数，tf 为词频，idf 为反文档频率，N 训练数据中文档的数量，$freq_{it_k}$ 为在文档 d_i 中包含特征 t_k 的个数。

3.2.2.2 特征提取

在已经得到的向量空间模型中，维度数十分巨大，对于分类而言，很多维度与类别关系不大，因而对分类意义不大，而且维度数巨大会严重降低分类器的分类速度。所以，需要提取出对分类意义重大的词汇，作为 VSM 的维度。

首先，需要去停用词，即去掉因出现频率过高，对分类效果不大的词汇。本实验利用文档频率法（document frequency）将在所有文档中出现最频繁的词建立停用词表，该过程主要利用 Java 程序实现，并选取频率最大的 30 个词作为停用词删除，停用词列表如下所示。取出停用词后所有文档中词的总数减少

了 17.79%。

停用词表：的，是，了，不，在，有，一，也，大，会，就，多，上，将，和，下，中，看，到，这，后，个，都，还，要，为，但，而，从，人

其次，取出文档中出现次数过低、不具代表性的词汇。由于频率过低的词汇在文档中对类别的代表性不强，利用它进行分类会使误差较大，经实验证明，取出词频数小于 3 的所有词时，分类的结果最好。经过去停用词和出现频率过低词两步后，VSM 的维度降低了 72.36%，初步达到了降低维度的效果，剩余 5811 个词汇，即 5811 个维度。

最后，特征提取。经过上述的简单处理，模型中的维度大大降低，但并非所有的特征都对分类是有用的，为分类器提供有效的输入数据，本实验再次对所剩余的维度进行进一步的筛选。本实验采用了实验效果较好信息增益（Information Gain，IG）,卡方（CHI），期望交叉熵（ECE）的特征选择方法，各方法均由 Java 程序编写。

3.2.3 分类与评估

3.2.3.1 分类器分类

支持向量机诞生至今，克服了“维数灾难”和“过学习”等传统困难，成为优秀的机器学习工具。本步骤选择当下应用相对普遍的支持向量机（SVM）、Naivebayes、决策树（J48）及 KNN 作为分类器，在 weka 数据挖掘软件里有 SVM 的算法实现，直接利用其提供的接口进行模型训练。

Weka 是一款机器学习、数据挖掘软件，是在 Java 环境下开发的开源软件，可以在其官方网站下载源码文件，该软件提供了图形界面，可以直接进行多算法执行，也可以利用 API 自己编写算。可以完成数据预处理、分类、聚类等一系列文本挖掘工作。Weka 能够处理 arff 及 cvs 等文件格式的数据。

3.2.3.2 分类结果评估

为了验证模型的有效性，本文采用 10 折交叉验证方法，即把实验数据分为 10 份，每次以其中 1 份作为预测数据，其他 9 份作为模型训练数据，共进行 10 次实验，对分类模型进行评估。由于本文最终利用股票相关评论来继续研究，所以在此仅考虑相关评论的召回率及准确率。

相关评论召回率如图 3-3 所示，召回率越高，说明有更多的相关评论被识别出来，损失的信息越少。对于特征提取算法来说，与 SVM 的结合是最优秀

的，且优于其他算法 10%以上。对于不同的分类算法来说，其最优的特征提取算法不尽相同。

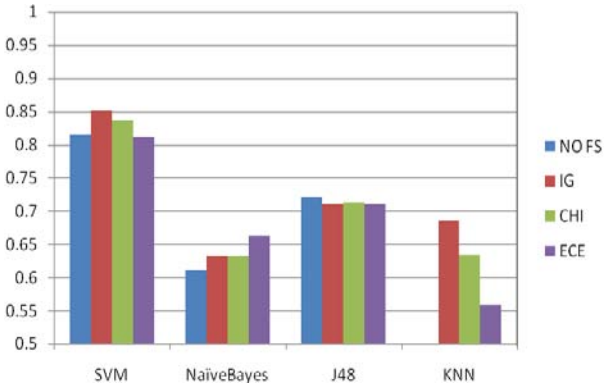

图 3-3 相关评论召回率

相关评论的准确率如图 3-4 所示，准确率越高，在所识别出来的相关评论中，真正与股票市场相关的评论比例较高，即信息质量较高。各分类算法都表现得相当优秀，Naivebayes 与 CHI 结合获得最高准确率， Naivebayes 结合 IG、SVM 结合 IG、SVM 结合 CHI 三个组合虽然表现略差，但相差不多。

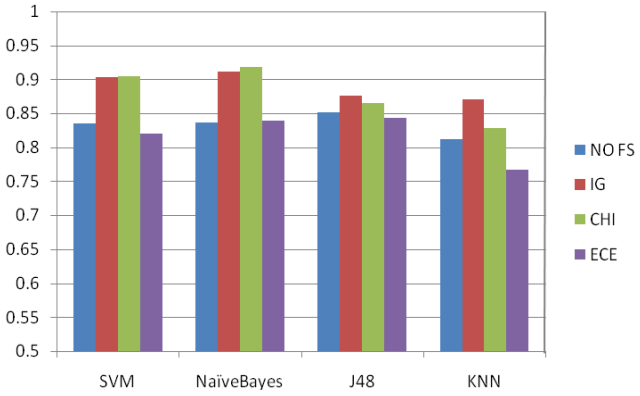

图 3-4 相关评论准确率

为综合考虑各组合的分类效果，本文主要对比的各组合的相关类别的 F 度量值如图 3-5 所示。IG 结合 SVM 略优于 CHI 与 SVM 组合，是最优组合，明

显优于剩余其它组合。综上召回率及准确率的对比，本文选择 IG 结合 SVM 作为最终分类组合。

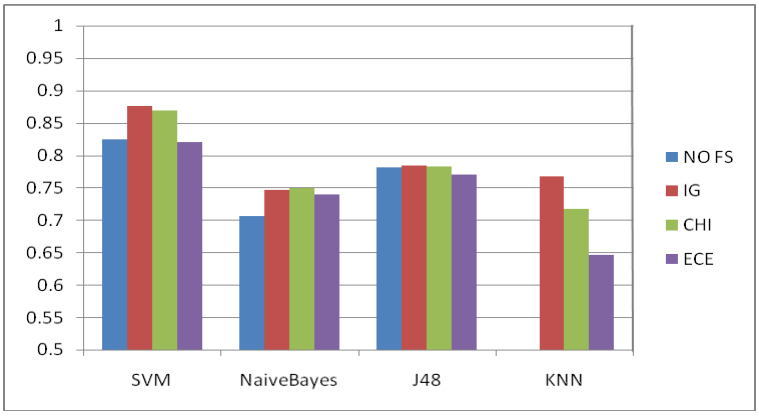

图 3-5 相关类别 F 值

以下为 IG 与 SVM 组合的详细分类结果如表 3-3 所示。

表 3-3 清除无关评论结果混淆矩阵

	被分为无关	被分为相关
实际为无关	305	58
实际为相关	95	542

无关类别及相关类别的查全率、查准率和 F-度量值如下表 3-4 所示。

表 3-4 清除无关评论结果评估

	Recall	Precision	F-measure
无关类别	0.8402	0.7625	0.8000
相关类别	0.8507	0.9030	0.8760

由分类结果可知，通过实验有 84.02%的无关信息被正确识别出来，基本清除了股票市场无关评论，同时 85.07%的相关评论被正确地归类，保证了评论信息的全面性。被分为相关信息的评论中，90.3%的评论为股票市场相关数据，保证了后续分析的数据正确性。

3.2.4 分类准确度影响因素分析

本小节主要讨论评论的长度对分类结果的影响，图 3-6 为 1000 条评论的长度的分布图，可见 50%的评论长度不足 100 字，40%的评论长度均与分布在 100 至 1000 字之间，剩余 10%的评论长度大于 1000，最大长度为 7464 字。

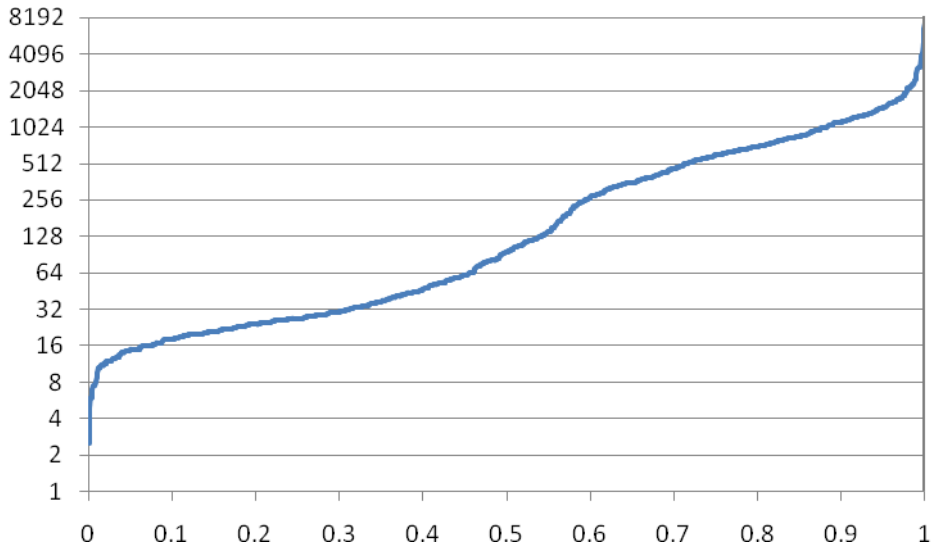

图 3-6 评论长度分布图

为了验证长度对评论分类结果的影响，本文进行对比做了如下实验，将长度大于某个值 X 的评论分为一组，长度小于 X 的评论分为另一组，分别进行上述实验。相关类别的准确率随评论长度变化如图 3-7 所示，当对所有评论长度大于 X 的评论集进行测试时，准确率随 X 的增大而逐渐增大，当 X 为 140 时，无关评论的召回率最高，之后有所下降，同时对所有有评论长度都小于 X 的评论集进行测试时，准确率随 X 的增加有所下降，X 为 140 时准确率最低。

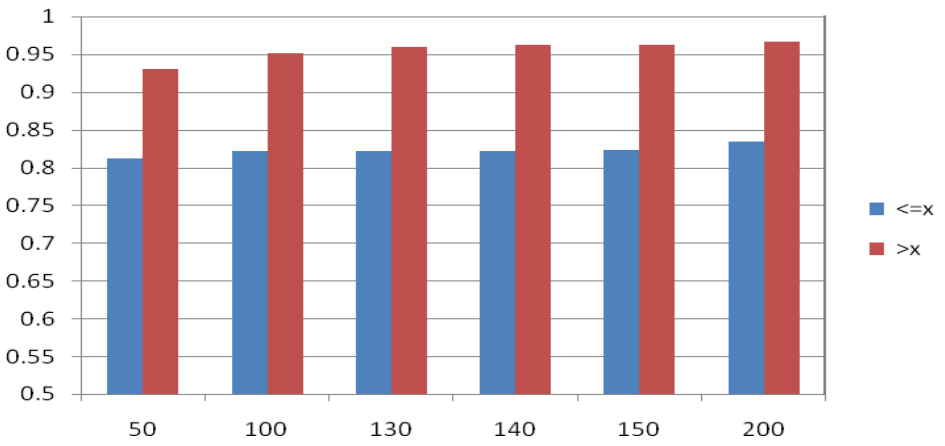

图 3-7 不同长度评论相关类别的准确率

相关评论的召回率随评论长度变化如图 3-8 所示，对于长度大于 50 字的所有评论的召回率就大于 92%,随着最小长度增加相关评论的召回率有所上升，最终基本维持在 96%左右。而对于长度小于 X 的所有评论，当 X 为 50 时召回率最小，为 69%，随 X 的增加有所上升，但变化不大。

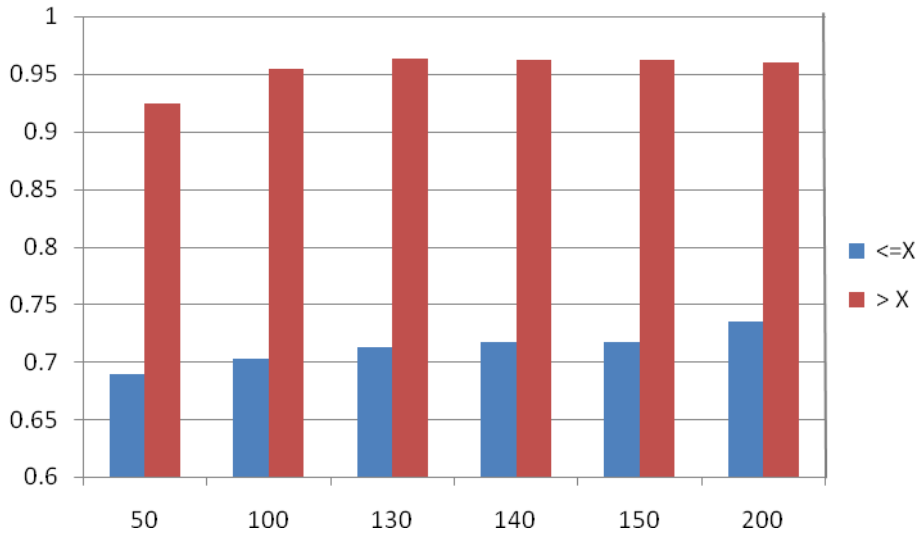

图 3-8 不同长度评论相关类别的召回率

由以上实验得出，评论的长度是影响分类准确度的关键因素，当评论较长时，无关评论的召回率可达到 87%以上，而相关评论的召回率可以达到 96%。

由此的出结论，受评论长度影响的主要是相关评论的召回率，对无关评论的影响不大。但无关数据清除的目的就是获取所有评论中的相关评论以进行下一步分析，所以，评论的长度对最终获取的相关评论的质量影响较大。

3.3 在线股评观点获取

通过对股票评论中的股票市场无关评论清除，可以获得股票市场相关评论。上述股票市场无关评论实验是基于大盘评论进行的，由于单只股票股价预测是本文的最主要研究内容，本部分实验采用中国联通的股票评论进行实验。运用以上方法，本文对中国联通的股票评论进行无关评论清除，得到分类准确率为83.07%，比大盘评论实验略低，这里主要是由于在上述选择的大盘评论，均为东方财富网已登录用户，评论更专业，评论平均长度更长，对无关评论的清除会有影响。

3.3.1 情感分类

在本文中认为，股票市场相关评论都含有对股票价格的情感倾向，并进一步对情感倾向进行识别。本文通过三种方法进行了观点识别并对比其结果，分别是 n 元文法（n-gram）、语义情感分析法（SO-A）、支持向量机（SMO）。

3.3.1.1 基于 SMO 的情感分类

SMO 由 Weka 提供算法实现，本文小节已对此软件及算法进行详细介绍，在此，不再赘述。其实现流程同无关数据清理流程，前文已详细介绍，在此不作过多说明。

3.3.1.2 基于语义情感分析法的情感分类

语义情感分析法（SO-A）算法在前一章也进行了详细介绍，本部分应用 Java 程序将算法实现，并对相关评论进行实验。以下对具体操作进行详细介绍。

步骤一，中文分词，前面介绍了在中分分词过程中引入了金融词库，但由于词库中的词汇未进行词性标注，SO-A 方法又需要对各个词进行词性标注，所以本算法不再引入金融词库，只使用 ICTCLAS 进行分词及词性标注。

步骤二，选取具有语义倾向的词语模式，由于没有现成的中文语义倾向词语模式，本文使用英文情感倾向的词性组合模式[45]进行替代，最常用的 5 个双词模式如表 3-5 所示。本步骤利用 Java 程序，结合正则表达式匹配出评论中包含以下五个双词模式的词组。

表 3-5 英文评论情感分析双词模式

	首词	尾词
模式 1	形容词	名词
模式 2	副词	形容词
模式 3	形容词	形容词
模式 4	名词	形容词
模式 5	副词	动词

步骤三，选择极端正面及负面代表词，本文主要研究评论者的情感倾向对股票价格的影响，评论者的情感倾向主要是认为股票会涨或会跌，所以本文采用“涨”和“跌”作为极端正面和极端负面的代表词。

步骤四，计算 SO 值，由于百度在国内的搜索引擎市场上所占的份额最大，其搜索结果具有一定的权威性，况且在娱乐生活搜索方面更远胜于其他搜索引擎，所以本文选择百度搜索结果作为衡量词语间相似度的数据来源。利用 Java 程序向搜索引擎发送请求，请求中包含极端词（涨或跌）及步骤二中所提取出的词组作为搜索关键词，并使用 Java 读取返回结果页，利用正则表达式匹配出返回的结果总页数，按照公式 2-3 所示计算每个词组的 SO 值。为提高效率，本步骤采用多线程方式，并将搜索过的词组及搜索结果存入数据库中，避免重复搜索浪费时间及网络资源。

步骤五，计算每个文档的平均 SO 值，即 SO-A 值，本文设定阈值为 0，SO-A 值大于 0 为看涨倾向，否则为看跌倾向。

3.3.1.3 基于 N-Gram 的情感分类

n-gram 有 lingpipe 实现。Lingpipe 是一款开源的自然语言处理软件包，主要包含以下模块：主题分类、命名实体识别、词性标注、句题检测聚类、字符语言建模、中文分词、情感分析（Sentiment Analysis）等。由于情感分析需要有领域情感词汇，但目前还没有较权威的中文金融类情感词库，本部分仍将股票评论的情感分析用文本分类方法解决。由于 lingpipe 自动实现文本分词，所以，只需将评论原文输入即可，省去了中文分词、特征提取、特征表示等步骤。

3.3.2 情感分类结果分析

SO-A 可以直接得出判断结果，n-gram 和 SMO 均采用十重交叉验证进行模

型评估。三种情感分析方法结果对比如表 3-6 所示

表 3-6 三种情感分析方法结果对比

		SO-A	SMO	Linpipe
涨	Recall	52.32%	83.46%	78.08%
	Precision	60.32%	81.25%	74.20%
	F-measure	56.04%	82.34%	76.09%
跌	Recall	57.47%	76.40%	66.21%
	Precision	49.37%	79.03%	71.0335%
	F-measure	53.11%	77.69%	68．54%
	Mic_Avg	54.62%	80.29%	72.79%

就分类结果表现来讲，SMO 要明显优于其它两种算法，语义分析方法表现最差。本文采用的语义分析方法的提出是应用于对具体产品评论的情感分析，由于股票评论与产品评论长度、评论内容、评论语言模式都有所不同，所以导致分类表现差。所以，在本文后续实验中采用 SMO，结合信息增益特征提取方法作为情感分类工具。

但从效率来讲，SO-A 方法避免了人工标定文本，节省大量的人力成本，对所有数据的处理都是相同的，适合大量数据的处理。Lingpipe 避免了中文分词、词性标注、特征提取等过程，虽然需要人工标定情感类别，但效率要优于 SMO。利用 SMO 实现过程需要大量的处理流程，需要人工参与的的最多，虽然分类结果最好，但是效率是最差的。而且随时间推移及数据量的增大，需要不断的扩充训练数据集。

3.4 本章小结

本章对情感分析的实验过程和结果进行了详细的描述，包括一下几个部分：（1）无关数据的清理过程及方法；（2）无关数据清理的实验结果分析，发现评论长度是影响实验结果的重要因素；（3）三种方法的情感分析实验、结果对比，其中 SMO 文本分类算法结合信息增益是最好的情感分析方法。实验取得了较好的结果，能够有效清除无关数据，并获得评论的情感倾向，为下一阶段的数据分析打下了基础。

第4章　基于网络论坛的股票市场分析

4.1 股评数据描述

在本文研究中,以 Alexa 中国的网站综合评定数据为标准,选择东方财富网、新浪、搜狐为在线股评的数据来源，并下载了三个网站的股评数据。为分析不同网站投资者发帖行为，本文以中国联通为例进行分析，下载了三个网站中国联通股吧板块的在线股评数据。由于搜狐只提供近半年的股评数据，综合其他两个网站，本部分选定 2011 年 10 月 1 日至 2012 年 3 月 6 日为数据来源区间，共 158 天，其中交易日为 100 天，非交易日为 58 天。下表为投资者发帖在各网站分布情况，可见，东方财富与搜狐网站用户发帖总量相差不大，新浪略少于其它两个网站。发帖行为大多发生在交易日中，非交易日发帖量非常少。在不同的交易时间，用户的发帖量也相差巨大。

表 4-1 各网站股票评论统计

	sina	eastmoney	sohu
总数	5198	7938	8215
交易日平均	49	71	74
非交易日平均	6	14	14
交易日最大值	445	560	561
交易日最小值	11	18	2

如下图所示，以小时为单位，对在线评论在一天中的分布情况进行了统计，描述了股票论坛参与者在一天中的活跃情况。在三个网站中，股评的分布情况是基本相同的，在每天的 10 点、14 点发帖量最大，其次是 11 点、13 点，再次是 9 点、15 点，即以 12 点为中心呈对称式分布。由此可以看出，股票论坛参与者在股票交易时间活跃程度较高，在其他时间活跃程度相对较低。在 21 点的时候又有了另一个发帖小高峰，这一高峰主要是取决于人们日常生活习惯。由于是休息时间，在 0 点至 7 点发帖数接近于 0。

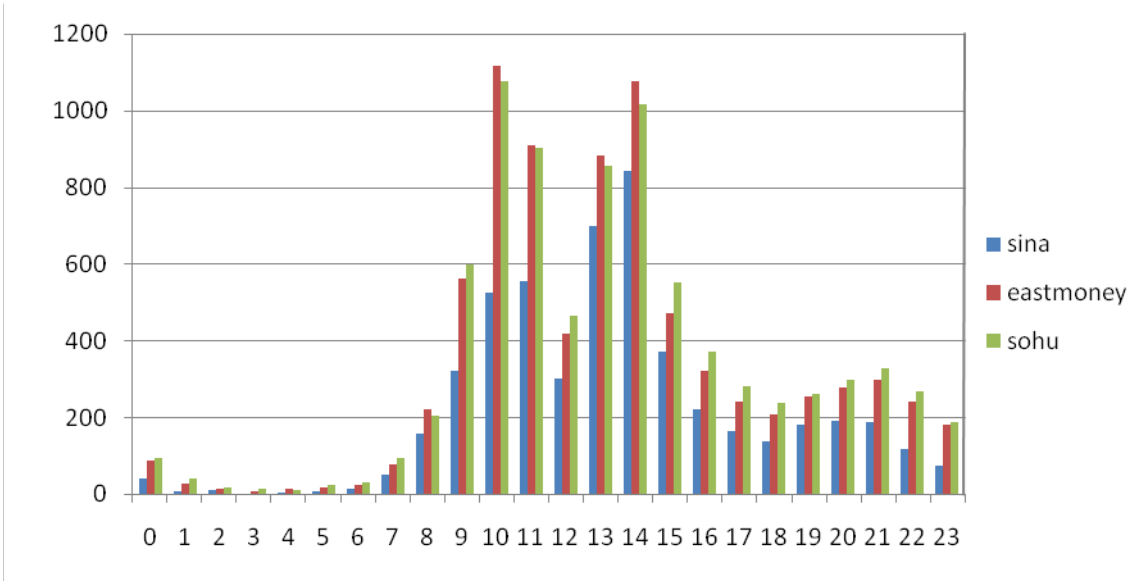

图 4-1 股票评论 24 小时分布图

如图 4-2 所示,统计了在线股评在一周之内的分布情况，在三个网站中，股评数量在双休日的数量都非常少，而在五个工作日中，股评数量相对较多且差异不大，周一到周四每天股评数量略有增加，到周五略有下降。

综上所述，三个网站中用户的发帖时间分布基本相同，由于三个网站在国内是用户分布最广的，可以推断，股票论坛参与者都有相似的发帖行为。用户的发帖行为并不随着所在网站的不同而不同。所以，在后续的分析中本文选择最具权威性的东方财富网作为股评数据来源。

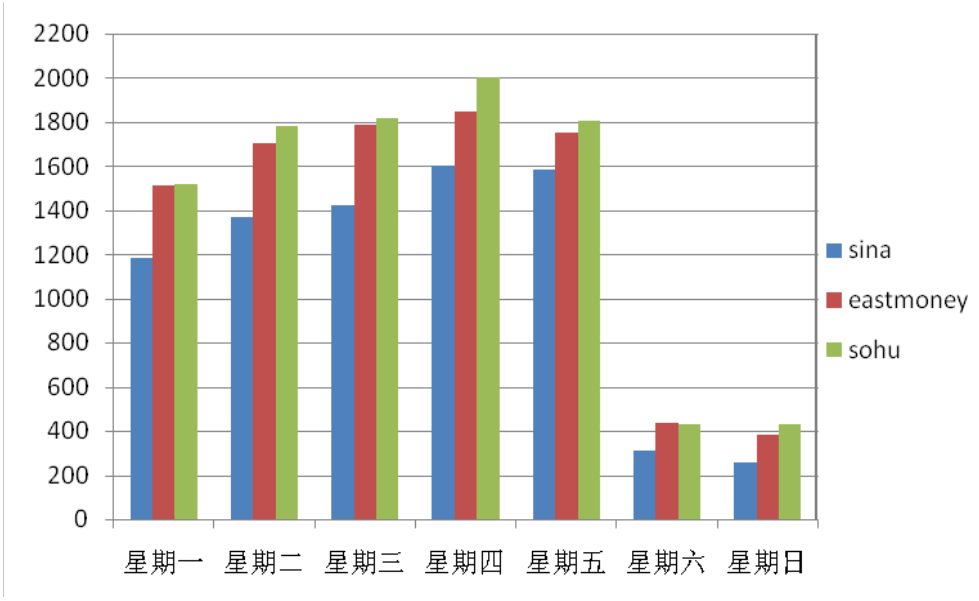

图 4-2 股票评论一周分布图

4.2 基于用户观点的股票市场分析

4.2.1 股价趋势预测变量

4.2.1.1 股价影响因素分析

股票价格的影响因素主要包括以下几个方面：

（1）宏观因素　主要包括经济周期、通货膨胀、货币政策、财政政策。经济周期是指经济要经历“复苏-繁荣-衰退-再复苏-再繁荣-再衰退”的循环过程，分为短周期、中周期、长周期，其中短周期为 3 至 4 年，本文主要研究股价的短期趋势预测，所以对于经济时期的影响不予考虑。

早期或温和的通货膨胀能够推动证券市场价格上涨，在通货膨胀时期，人们更倾向于购买实物商品来保值而不是股票，从而导致股票价格下跌。

货币政策的主要工具包括法定存款准备金及利率等。宽松的货币政策，即降低存款准备金率及利率，为公司营造了良好的融资环境，有利于证券市场的发展，促进股票价格上涨。而紧缩的货币政策从总体上会抑制股票价格的上扬。

财政政策主要包括国家预算、随手国债、财政补贴等政策。宽松的财政政策是指国家通过减免税负扶持经济、产业发展，股利民间资本进入证券市场等，总体上会促进证券价格的上涨。与此相反，紧缩的货币政策会使股票价格下跌。

（2）行业因素　行业是影响公司投资价值的主要因素之一，主要包括行业市场类型、行业生命周期、影响行业兴衰因素以及行业的收益性与成长性分析。

（3）公司因素　股票自身价值是决定股价最基本的因素，而这主要取决于发行公司的经营业绩、资信水平以及连带而来的股息红利派发状况、发展前景、股票预期收益水平等。

（4）市场因素　股票在证券市场上进行交易，必然受到股市变化的综合影响。

（5）心理因素　投资者的非理性或有限理性心理会对股票市场产生影响，过度自信、羊群效应等现象在现代股票市场中十分常见，对股票市场有着不可忽视的影响作用。投资者中包含个人投资者和机构投资者两部分.

结合以上分析，本文总结了影响股票价格的各个因素，并在以下的股价预测模型中将各个影响因素作为实际的预测变量。

本文将利率作为宏观经济发展形势的一个表征变量。著名的经济学家凯恩斯(Keynes)曾说过: 经济体系中的任何一个因素都与利率有一定关系。作为经济学家探索和研究的重要领域，政府调控市场的政策工具，利率是一个国家最重要的宏观经济变量。上海银行间同业拆放利率（Shanghai Interbank Offered Rate，简称 Shibor）是中国货币市场上的基准利率, 是我国利率体系中的主导性利率, 是中央银行金融宏观调控的基础性指标。所以，在本文中以 Shibor 为作为宏观经济变量，度量其对股票市场的影响。Shibor 主要有隔夜、1 周、2 周、1 个月、3 个月、6 个月、9 个月及 1 年期等几个品种，本文主要研究短期内的股票市场预测，所以，选择隔夜的 shibor 作为自变量。

将电信业务行业指数作为中国联通所在通信行业的发展变量。该指数代表电信行业领军企业的股票表现情况，对整个通信行业的发展有表征作用，所以，本文将该指数作为影响股价的行业因素变量。

公司基本情况在短时间内是维持稳定的，会通过公司公告将公司的重要新闻信息公之于众，所以本文采用公司公告作为公司变量产生变化的标识。公告数据来自东方财富网的公司公告板块，在所选定时间段内共有 89 条公告，分布在 61 天当中。包含定期报告（13）、股东大会（8）、股权变动（2）、股权交易（4）、可转债（13）、权益分派（2）、上市公司制度（2）、业绩预告、快报（1）、其他重大事项（40）。在此，本文只研究发公告是否对股票价格产生影响，如何

影响，但不考虑公告的具体内容。

投资者心理是本文主要的研究变量，投资者主要分为机构投资者和个人投资者。个人投资者情绪采用情感指数标识（在下小节中计算），机构投资者情绪利用机构评级。机构评级(Agency ratings, AR)：投资者主要包括两部分，个人投资者和机构投资者，本文中以股票论坛中的评论所表达的情感倾向作为个人投资者的情感倾向，机构投资者的情感倾向使用机构评级来表示。中国联通股票的评级来自新浪财经,在所选时间段内共有来自53家机构的728条评级数据，买入、持有、中性、减持、卖出五个等级分别有182、407、125、5、9条评论，分布在328天中。在处理评级数据时将买入、持有、中性、减持、卖出五个等级分别赋予-2、-1、0、1、2五个得分，并一天计算每日的平均得分作为当天的机构评分。

4.2.1.2 影响股价指标及股票市场表现指标选取

本研究以短期内股票市场为研究对象，借鉴了相关研究中测量股票市场及股票评论变量，以下为在本文中涉及的变量，以天为单位的计算方法。

（1）股票价格　本文采用股票在市场上的交易价格作为研究对象,包括每天收盘价、开盘价、最高价、最低价。

（2）上证指数　其样本股包括上证A股及B股的多有股票,。从总体上反映了上海证券交易所上市股票价格的变动情况，自1991年7月15日起正式发布。1991年启用，基期1990年12月19日，基点为100。

（3）上证电信业务行业指数 在上海证券电信行业中选择规模大、流动性好的股票组成样本股， 以反映上海证券市场该行业行业公司股票的整体表现，电信业务行业指数的成分股包括中国联通、大唐电信、永鼎通信等九只股票。2009年1月启用，以2003年12月31日为基期，基点为1000。

（2）股票交易额 股票在交易日内成交金额。

（3）换手率　指日内股票转手买卖的频率，日内股票成交量比发行总股数。

（4）波动率　每天股票的最高价与最低价之差除以开盘价与收盘价的平均值。

（5）收益率　第t天的收盘价与第t-1天的收盘价之差除以t-1天收盘价的百分比。

（6）机构评级 来自新浪财经-数据中心，包括买入、持有、中性、减持、

卖出等五个评级。由此变量本文衍生出两个变量，分别进行分析。包括机构评分和机构评级数，机构评分为每日各机构打分的平均值，机构评级数位每天进行评级的个数。

（7）公司公告 来自东方财富网特色数据板块，包括定期公告、可转债、股权交易等内容。此变量也由两个变量度量，分别为是否有公告和公告数，是否有公告某天有公告发表为 1，没有公告发表为 0，公告数表示某天公司发表的公告数量。

（8）公司新闻 本文着重研究了单只股票的股价影响因素及预测，由于关于公司的新闻包含公司的发展情况，是投资者在投资过程中参考的重要信息，本文将该变量作为控制变量纳入股票价格预测模型中。

（9）在线股评数量 股票论坛中每天发表帖子的数量。具体统计过程中，以 t-1 天 15 点至 t 天 15 点之间发表的帖子数量作为第 t 天在线股评数量进行计算。

（10）情感指数 在本文中该指数有每天中各评论的具体情感倾向（看涨或看跌）计算而得。借鉴已有的文献，本文应用三个情感指数来度量每日网络论坛中用户的情感倾向。

简单情感指数（Smple Sentiment index，SSI）：

$$SSI = M^{Bull} - M^{Bear} \tag{4-1}$$

看涨指数（Bullishness index，BI）[55]：

$$BI = \ln\left[\frac{1+M^{Bull}}{1+M^{Bear}}\right] \tag{4-2}$$

情感差异指数（Distinguish index sentiment，DIS）[32]表示在股票论坛中各投资者的情感差异程度：

$$DIS = \left|1-\left|\frac{M^{Bull}-M^{Bear}}{M^{Bull}+M^{Bear}}\right|\right| \tag{4-3}$$

其中，M^{Bull} 为看涨评论的数量，M^{Bear} 为看跌评论的数量

4.2.1.3 投资者情绪与股票表现变量相关分析

由于我国股市发展时间相对较短，与国外成熟的证券市场相比，出于投机心理的短线交易者较多。由于我国股市的这一特点，本文选择股票评论与股票

表现在短期内的相互影响作为研究对象。

中国联合网络通信集团有限公司，简称“中国联通”，2009 年 1 月由原中国网通和原中国联通组建而成，是中国唯一一家在纽约、香港、上海三地同时上市的电信运营企业，连续多年入选“世界 500 强企业”，是电信行业龙头企业，是绩优大盘股，具有代表性作用，本文选中国联通作为单只股票进行研究。

利用 SPSS 软件对各变量进行相关分析, Pearson 相关系数表如表 4-2 所示。BI 与 SSI 都是描述股票论坛中投资者的总体情感倾向的，BI 指数仅与收盘价、收益率显著相关，但相关程度不强，SSI 与此二变量也显著相关，且相关程度强于 BI。此外，SSI 还与交易额、换手率相关，但相关性都不强。DIS 是描述股票论坛中投资者之间的情感差异的，仅与交易额、换手率呈弱正向相关关系。股评数量反应投资者每天发表的股评总数量与交易额、换手率有较强的相关关系。机构投资者情绪机构评分，与收盘价、交易额、换手率不相关，与波动率、收益率有弱相关性。综上所述，本文所采用的投机者情感指数与股票市场表现具有一定的相关性，本文采用的投资者情绪变量是有意义的。

表 4-2 相关系数表

	收盘价	交易额	换手率	波动率	收益率
BI	.117*	-.039	-.053	-.058	.232**
SSI	.186**	.135**	.110*	.036	.349**
DIS	-.032	.135**	.145**	.047	-.073
股评数量	.096*	.569**	.602**	.166**	-.034
机构评分	.057	.049	.041	.087*	.145**

**. 在 .01 水平（双侧）上显著相关。

*. 在 你 0.05 水平（双侧）上显著相关。

如图 4-3 为股票价格与 BI、SSI 两情感指数的趋势图，可以看出，两情感指数的走向与股票收盘价基本相同，进一步证明其存在一定的相关性。

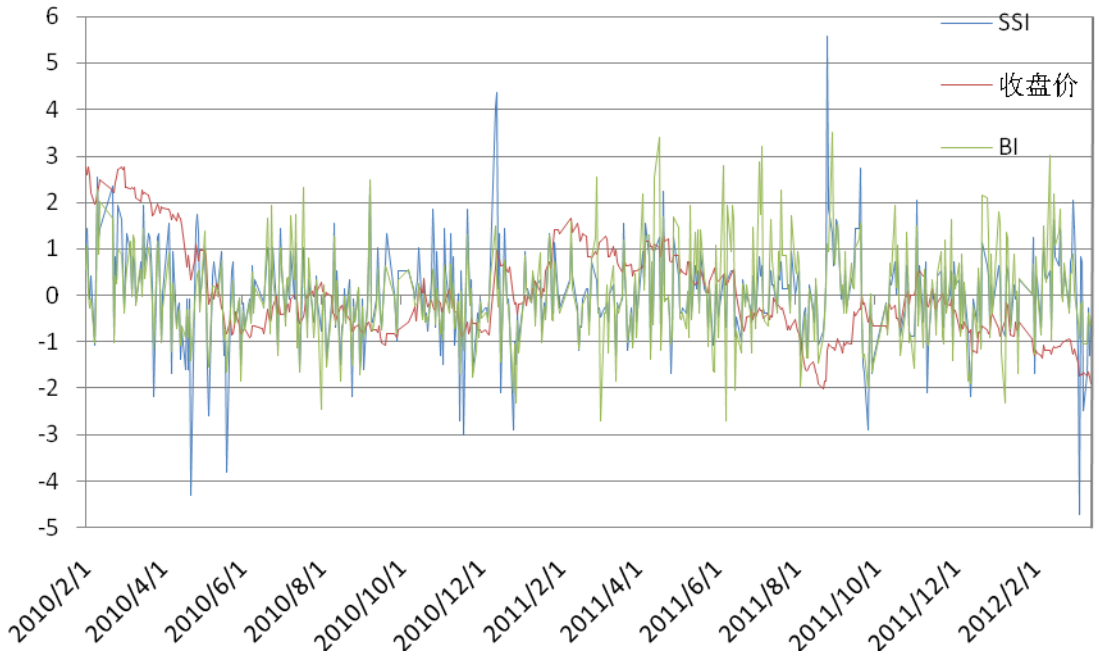

图 4-3 情感趋势图

4.2.2 股价预测模型分析

4.2.2.1 多元线性回归分析模型假设

运用多元线性模型进行分析，必须满足以下基本假设[56]：

假设 1：各解释变量是随机的，且相互之间互不相关，即不存在多重共线性。

假设 2：随机干扰项均值为 0，同方差且不存在序列相关。

假设 3：解释变量与随机干扰项不相关

假设 4：随机干扰项满足正态分布。

假设 5：当样本量趋于无穷时，个解释变量的方差趋于有界常数。

假设 6：回归模型的设定是正确的。

4.2.2.2 多元回归预测模型

综合股票价格的影响因素分析，本文建立股票价格回归模型，以股票每日收盘价为因变量，其它为自变量，自变量有前五日股票收盘价（price）、行业指数（TSII）、上证综合指数（Market），简单情感指数（SSI），看涨指数（BI），

情感差异指数（DIS），股评数量（Comment），机构评级（Agency），公司公告（Announcements），新闻变量（News），利率（shibor）等变量。模型如式 4-4 所示。

$$\begin{aligned} Price_t = \beta_0 &+ \sum_{i=1}^{5}\sum_{j=1}^{5} Price_{i,t-j} + \sum_{i=6}^{10}\sum_{j=1}^{5} Market_{i,t-j} + \sum_{i=11}^{15}\sum_{j=1}^{5} TSII_{i,t-j} + \sum_{i=16}^{20}\sum_{j=1}^{5} SSI_{i,t-j} \\ &+ \sum_{i=21}^{25}\sum_{j=1}^{5} BI_{i,t-j} + \sum_{i=26}^{30}\sum_{j=1}^{5} DIS_{i,t-j} + \sum_{i=31}^{35}\sum_{j=1}^{5} Comment_{i,t-j} + \sum_{i=36}^{40}\sum_{j=1}^{5} Agency_{i,t-j} + \\ &\sum_{i=41}^{45}\sum_{j=1}^{5} News_{i,t-j} + \sum_{i=46}^{50}\sum_{j=1}^{5} \text{Announcements}_{i,t-j} + \sum_{i=51}^{55}\sum_{j=1}^{5} Shibor_{i,t-j} + \varepsilon_t \end{aligned} \quad (4\text{-}4)$$

本部分利用中国联通 2010 年 2 月 1 日至 2012 年 3 月 16 日的数据进行实验，其中包括 511 个交易日。

首先，模型将模型中变量与收盘价格进行相关分析，部分结果如表 4-4 所示。

表 4-3 收盘价格与各预测变量 Person 相关系数表

	滞后 1 天收盘价	滞后 2 天收盘价	滞后 3 天收盘价数	滞后 4 天收盘价数
Pearson 相关性	.981**	.961**	.943**	929**
显著性（双侧）	.000	.000	.000	.000
N	511	511	511	511
	滞后 5 天收盘价数	公告数量	机构评级分	滞后上证指数
Pearson 相关性	-.917**	.069	.057	.694**
显著性（双侧）	.000	.125	.210	.000
N	511	511	511	511
	滞后行业指数	滞后一天新闻数量	新闻数量	
Pearson 相关性	.558**	-.071	.057	
显著性（双侧）	.000	.110	.210	
N	511	511	511	

**. 在 .01 水平（双侧）上显著相关。 *. 在 .05 水平（双侧）上显著相关。

可以看到股票收盘价与滞后几天的收盘价呈显著正相关，且相关性非常强说明股票的价格受之前价格影响较大。收盘价与 Shibor、滞后上证指数、滞后

行业指数相关性较强，在 0.01 的置信水平下显著。从公司公告、机构评论及新闻变量均与收盘价不存在显著相关。

将以上变量均作为自变量对股票收盘价进行逐步回归，利用 SPSS 运算结果如模型汇总如表 4-4 所示。先后将七个自变量纳入回归方程，分别为滞后 1 天收盘价, 简单情感指数, 滞后 5 天收盘价, 机构评级分， 滞后一天新闻数，滞后三天行业指数。回归模型最终调整 R 方为 0.969，说明本文建立的模型能够对股票收盘价的大部分变化进行解释。DW 值为 1.886，说明该模型不存在序列相关，即满足多元线性回归的基本假设之一，模型的随机干扰项互相独立。

表 4-4 模型汇总

模型	R	R 平方	调整 R 平方	标准估计的误差	Durbin-Watson
1	.981[a]	.962	.962	.09204	
2	.983[b]	.967	.967	.08593	
3	.984[c]	.968	.968	.08517	
4	.984[d]	.968	.968	.08478	
5	.984[e]	.969	.968	.08442	
6	.984[f]	.969	.969	.08386	
7	.985[g]	.969	.969	.08361	1.886

a. 预测变量: (常量), 滞后 1 天收盘价。

b. 预测变量: (常量), 滞后 1 天收盘价, 简单情感指数。

c. 预测变量: (常量), 滞后 1 天收盘价, 简单情感指数, 滞后 5 天收盘价。

d. 预测变量: (常量), 滞后 1 天收盘价, 简单情感指数, 滞后 5 天收盘价, 机构评级分。

e. 预测变量: (常量), 滞后 1 天收盘价, 简单情感指数, 滞后 5 天收盘价, 机构评级分, 滞后一天新闻数。

f. 预测变量: (常量), 滞后 1 天收盘价, 简单情感指数, 滞后 5 天收盘价, 机构评级分, 滞后一天新闻数, 滞后三天行业指数。

g. 预测变量: (常量), 滞后 1 天收盘价, 简单情感指数, 滞后 5 天收盘价, 机构评级分, 滞后一天新闻数, 滞后三天行业指数, 滞后四天新闻数。

h. 因变量: 收盘价

表 4-5 为方差分析表，通过 F 检验说明回归方程是显著的。模型可以看出，对收盘价解释能力最强的变量是滞后一天收盘价，这是由于股价价格有记忆功能，前一天的收盘价格是投资者交易的最重要参考指标，而滞后五天收盘价也被纳入模型，对模型的解释能力有所提高，充分说明了股票价格具有记忆功能。简单情感指数对股票收盘价解释能力仅次于滞后收盘价，而其他两个情感指数及股评数量在逐步回归过程均为纳入模型，说明简单情感指数对股票收盘价的解释能力最强，其他变量的解释能力较弱。在本文中机构评分被作为机构投资者情绪变量引入，在继滞后收盘价，简单情感指数之后引入模型，对模型的解释能力有所提高。滞后一天和三天的新闻数量也对模型有显著的解释作用，说明公司的新闻能够解释后续的股票价格变化，对股票投资者有参考作用。滞后三天电信业务行业指数也对自变量有解释作用，说明股票所在的行业的发展对后续股票价格有影响作用。

本模型也引入了滞后一天至五天的简单情感指数及机构评级数作为自变量，但均因为不显著而未被保留下来，无论是机头投资者情绪还是个人投资者情绪，都直接影响股票价格，不存在一个延迟的作用。而滞后四天的新闻有解释能力，新闻对股票价格的影响有持续作用。

表 4-5 Anova

模型		平方和	df	均方	F	Sig.
1	回归	110.088	1	110.088	12994.342	.000a
	残差	4.312	509	.008		
	总计	114.400	510			
2	回归	110.648	2	55.324	7491.605	.000b
	残差	3.751	508	.007		
	总计	114.400	510			
3	回归	110.722	3	36.907	5087.945	.000c
	残差	3.678	507	.007		
	总计	114.400	510			
4	回归	110.762	4	27.691	3852.082	.000d
	残差	3.637	506	.007		
	总计	114.400	510			

表 4-5 （续表）

模型		平方和	df	均方	F	Sig.
5	回归	110.801	5	22.160	3109.626	.000e
	残差	3.599	505	.007		
	总计	114.400	510			
6	回归	110.855	6	18.476	2627.147	.000f
	残差	3.544	504	.007		
	总计	114.400	510			
7	回归	110.883	7	15.840	2265.900	.000g
	残差	3.516	503	.007		
	总计	114.400	510			

由 4-4 的 P-P 图可以看出：散点几乎随机地围绕在角分线上，因此，误差近似服从正态分布。由残差的散点图 4-5 所示，可以看出散点几乎随机且均匀地分布在零线上下方，因此，误差的等方差性和独立性基本成立。

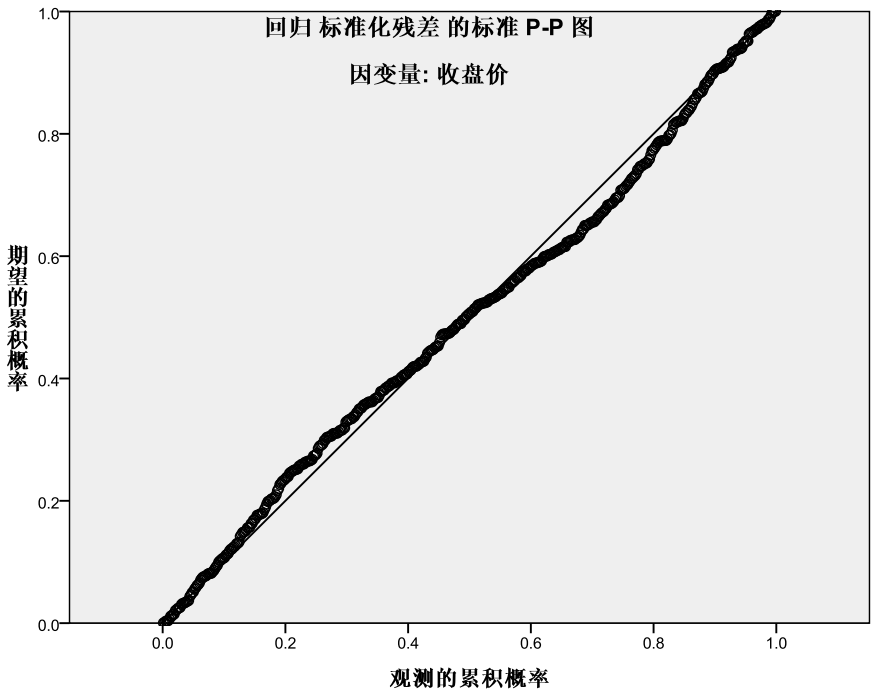

图 4-4 P-P 图

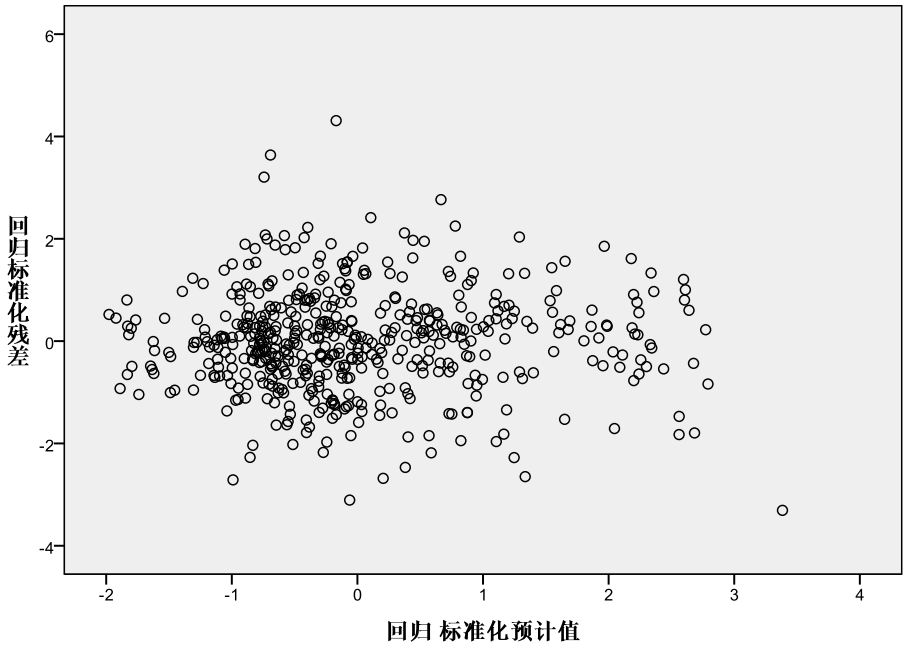

图 4-5 残差图

如表 4-6 为模型的系数表，通过 t 检验可知，模型的前五个系数系数在 0.01 的置信水平下是显著地，另外两个变量在 0.05 的置信水平下是显著的。各变量的 VIF 都小于 10，说明模型不存在多重共线性。除滞后一天新闻数外,其他变量均与股票价格正相关，即前一天出现新闻数越多，对股票价格负面影响越大，而在前四天有较多的新闻，对股票价格的影响是正向的。投资者的情感指数对股票价格的影响也是正向的，说明投资者情感越积极，对股票价格有向上的推动作用。滞后一天收盘价及滞后五天收盘价对股票收盘价有正向影响，说明投资者在股票价格较高，对将来股票就有正向预期，在股票价格较低时，则相反。而滞后一天收盘价的系数要大于滞后五天收盘价的系数，说明股票价格受前一天的收盘价格影响较大。通信行业指数对股票价格的影响要有一定的延迟，并且对中国联通的股价有推动作用，说明股票价格受行业整体发展的影响，但影响有一定的延迟，行业的变化不能马上表现在股票价格上面。

通过以上实验，得到股票价格预测模型如下所示

$$\begin{aligned} Price_t = {} & 0.161 + 0.894 * Price_{t-1} + 0.004 * SSI_t + \\ & 0.059 * Price_{t-5} + 0.006 * Agency_t - 0.002 * News_{t-1} \\ & + 0.00003578 * TSII_{t-3} + 0.001 * News_{t-4} \end{aligned} \tag{4-5}$$

表 4-6 最终回归模型系数表

	非标准化系数		标准系数			共线性统计量	
	B	标准误差		t	Sig.	容差	VIF
(常量)	.161	.044		3.640	.000		
滞后 1 天收盘价	.894	.022	.901	40.270	.000	.122	8.190
简单情感指数	.004	.000	.075	9.293	.000	.937	1.067
滞后 5 天收盘价	.059	.021	.062	2.823	.005	.129	7.776
机构评级分	.006	.002	.021	2.624	.009	.977	1.024
滞后 1 天新闻数	-.002	.001	-.026	-3.160	.002	.928	1.078
滞后 3 天行业指数	3.578E-5	.000	.023	2.357	.019	.631	1.584
滞后 4 天新闻数	.001	.001	.016	2.005	.046	.926	1.079

a. 因变量: 收盘价

4.2.2.3 回归模型预测结果检验

为验证模型的显著性，根据预测模型设计规则[57]，将公式 4-5 所示模型进行在模型样本数据以外的数据进行验证，采用中国联通 2012 年 3 月 19 日至 2012 年 4 月 27 日的数据对模型进行验证，其中包括交易日 26 个。预测样本数据及结果如表 4-7 所示，平均预测准确率为 98.5%。

预测准确率的计算方法如下：

$$Precision_t = \left(1 - \frac{\left|\widehat{price_t} - price_t\right|}{price_t}\right) * 100\% \tag{4-6}$$

其中，$\widehat{price_t}$ 为预测 t 日股票收盘价格，$price_t$ 为 t 日实际股票收盘价格，$Precision_t$ 为 t 日股票价格预测准确率。

表 4-7 预测检验

收盘价	简单情感指数	滞后一天收盘价	滞后五天收盘价	滞后一天新闻数	滞后四天新闻数	机构评分	滞后三天行业指数	预测价格	预测准确率
4.55	39	4.58	4.62	8	15	1	1545.48	4.744397	0.957275
4.48	-20	4.55	4.65	23	9	1.4	1531.76	4.449256	0.993138
4.48	-5	4.48	4.6	2	8	1	1551.47	4.483032	0.999323
4.54	38	4.48	4.53	22	8	1	1561.96	4.611277	0.9843
4.48	15	4.54	4.58	12	23	0.85	1525.2	4.608629	0.971288
4.47	5	4.48	4.55	14	2	1.33	1522.35	4.491040	0.995293
4.48	0	4.47	4.48	24	22	1	1528.17	4.456178	0.994683
4.33	-20	4.48	4.48	17	12	1	1507.9	4.388393	0.986514
4.2	-34	4.33	4.54	16	14	0	1506.8	4.199793	0.999951
4.23	-39	4.2	4.48	8	24	2	1512.55	4.098239	0.968851
4.26	15	4.23	4.47	8	17	0	1453.43	4.319354	0.986067
4.21	-7	4.26	4.48	5	16	0	1423.65	4.262698	0.987483
4.16	21	4.21	4.33	6	8	1	1425.07	4.317199	0.962212
4.23	-30	4.16	4.2	13	8	1	1450.47	4.047738	0.956912
4.21	1	4.23	4.23	6	5	2	1453.14	4.253183	0.989743

表 4-7（续表）

收盘价	简单情感指数	滞后一天收盘价	滞后五天收盘价	滞后一天新闻数	滞后四天新闻数	机构评分	滞后三天行业指数	预测价格	预测准确率
4.25	-1	4.21	4.26	11	6	0	1432.26	4.207326	0.989959
4.29	-2	4.25	4.21	4	13	0	1448.32	4.257711	0.992473
4.27	27	4.29	4.16	9	6	0	1447.09	4.389477	0.972019
4.24	-9	4.27	4.23	14	11	1	1468.19	4.233482	0.998463
4.28	-2	4.24	4.21	5	4	0	1483.18	4.239018	0.990425
4.26	5	4.28	4.25	13	9	2	1489.55	4.306366	0.989116
4.35	24	4.26	4.29	8	14	1.2	1480.32	4.376716	0.993858
4.41	24	4.35	4.27	5	5	0	1505.19	4.446686	0.991681
4.35	9	4.41	4.24	11	13	0	1497.25	4.434272	0.980627
4.37	5	4.35	4.28	6	8	0	1518.26	4.372743	0.999372
4.38	42	4.37	4.26	16	5	0.8	1533.96	4.519805	0.968081

4.2.3 股票价格走向影响因素分析

股票价格走向是投资者投资过程中关心的重要信息，将股票价格走向作为分类问题进行研究，即股票价格上涨为一类，股票价格下跌为一类，而连续两日收盘价相同，价格走向为平的情况较少，且股票价格走平，对投资者来说投资意义不大，与股票价格下跌相差无异，所以本文将其视作“跌”类。

为充分考虑影响股票价格走向影响因素，本文将以下变量纳入实验，过去 10 天的情感指数（看涨指数、简单情感指数、情感差异指数）、收盘价格、评论总数、shibor、机构评分、公告数量、新闻数量作为预测变量。另外，考虑到股票价格走势是一个相对值，即相对于前一天的收盘价的涨跌方向，可能受到过去价格走势、行业为指数走势及大盘走势的影响，本文将过去 10 天股票收盘价、交易量、电信业务行业指数、上证综指涨跌额作为预测变量。

通过卡方和信息增益的方法对各变量进行特征提取，即提取对分类有意义的变量。结果简单情感指数、看涨指数、滞后一天新闻数是影响股票价格走向的影响因素，其他变量的加入均未能给股票价格走势的预测带来新的信息，即对预测股票价格走势无意义。

通过信息增益方法计算，所有变量中只有看涨指数、简单情感指数、滞后一天新闻数三个变量的信息增益值大于 0，分别为 0.0285、0.0273、0.0235，其他变量的信息增益值均为 0，即对判别股票涨跌无效。利用卡方的方法与信息增益结果相同，得到的三个变量也为看涨指数、简单情感指数、滞后一天新闻数，卡方值分别为 19.936、19.0272、16.3648。

继续基于上述得到的三个变量，利用 SMO 方法分类，进行股票走向预测，预测准确率为 61.06%。

4.3 通讯行业分析

为进一步研究不同股票在线股评对股票市场影响，本文将东方财富网通信行业数据作为数据来源进行研究，从截面数据角度分析在线股评的影响。

4.3.1 股评变量与股票表现相关分析

截止至，2012 年 3 月，通讯行业共有股票 71 只，其中包括上证 A 股 18 支，深证 A 股 23 支，创业板 17 支，中小板 23 支。其中各股票上市时间差异较大，本文下载了 2010 年 11 月 1 日至 2012 年 3 月 16 日来自东方财富网的股吧板块的股评数据，并利用第三章的情感分析方法分析了每条评论多包涵的投资者情绪（看涨、看跌）。在分析过程中，剔除了上市时间晚于 2010 年 11 月 1 日的股票，剩余 60 只股票。剔除了 60 支股票的非交易日股票评论,剩余 282366 条评论。其中，中国联通最为活跃，平均日评论量为 73 条，最不活跃的股票是日海创业及华星创业，平均日评论量为 2 条。

由于在上一小节实验中，简单情感指数表现最好，在本小节将对通讯行业各支股票的收盘价、交易量分别于简单情感指数及股评数量进行了相关分析，结果如表 4-9 所示。其中，有 20 支股票的收盘价格与情感指数显著相关，但相关性不强，其中 4 支股票的相关系数是负的。38 只股票的成交量与简单情感指数显著相关，相关系数最大的是奥维通信为 0.51，这 38 支股票中，仅有一只股票的相关系数是负的，为负的 0.38。有 14 支股票的收盘价、成交量均与简单情感指数相关，且相关系数最大的均是奥维通信。可以明显看出简单相关系数与股票成交量的相关性要强于与股票价格的相关性。

53 支股票的收盘价均与股评数量相关，拥有最大相关系数的是东方电信，为 0.671，其中 13 只股票的相关系数为负的。59 支股票的成交量均与股评数量

相关，仅梅泰诺一支股票相关性不显著，北纬通信的相关系数更是高达 0.886，所有成交量与股评数量的相关系数均为正，即股评数越多，关注该股票的投资者越多，交易越活跃，股票的成交量就越高。

综合来看，股评数量与股票价格、成交量的相关性都要强于简单情感指数。而二者对成交量的相关系数几乎都是正的，即，投资者情绪越积极，股评的数量越多，股票的成交量就越大。而对于收盘价，简单情感指数、股评变量与其相关系数都有正有负。

表 4-8 相关分析表

股票	平均日股评数量	简单情感指数与收盘价	简单情感指数与成交量	股评数量与收盘价	股评数量与成交量
奥维通信	4	.400**	.510**	.641**	.749**
华平股份	7	.325**	.390**	.455**	.495**
高新兴	6	.264**	.241**	.450**	.428**
中创信测	10	.243**	.471**	.333**	.617**
烽火电子	19	-.198**	0.001	.254**	.562**
中国联通	73	.186**	.176**	-.215**	.649**
特发信息	6	-.170**	0.071	.194**	.383**
数码视讯	17	.160**	.151**	.359**	.154**
长江通信	24	.149**	.231**	.377**	.548**
ST 波导	24	.148**	.136*	.591**	.758**
大唐电信	32	.145*	.243**	.470**	.526**
中卫国脉	26	-.141*	0.059	.340**	.578**
大富科技	14	.141*	.230**	-0.011	.545**
烽火通信	10	.135*	0.106	-.196**	.319**
*ST 科健	7	.127*	.157*	.154*	.545**
*ST 星美	13	.126*	0.004	.160**	.490**
东信和平	11	.117*	.313**	0.06	.580**
梅泰诺	6	-.116*	0.063	-.375**	0.013
键桥通讯	7	.032**	.124**	-.093**	.612**
合众思壮	10	.031**	.124**	-.100**	.612**

表 4-8（续表）

股票	平均日股评数量	简单情感指数与收盘价	简单情感指数与成交量	股评数量与收盘价	股评数量与成交量
武汉凡谷	13	0.098	.172**	.135*	.359**
三元达	10	0.095	.132*	-0.008	.625**
光迅科技	3	0.092	.202**	.122*	.663**
恒信移动	14	0.083	.208**	-.151**	.762**
中天科技	28	0.081	0.082	.152**	.448**
电子城	12	0.062	.136*	.414**	.389**
中兴通讯	11	0.061	-.380**	-.481**	.514**
国脉科技	9	0.06	.187**	-.154**	.509**
盛路通信	5	0.06	.306**	0.033	.492**
航天通信	18	0.036	0.07	.395**	.667**
高鸿股份	25	0.035	-0.064	.170**	.672**
深信泰丰	10	0.031	.155**	.290**	.782**
同洲电子	16	0.026	0.08	.344**	.722**
振华科技	9	0.025	.137*	.387**	.446**
亨通光电	13	0.022	.303**	0.027	.643**
北纬通信	11	0.02	.266**	-.229**	.886**
日海通讯	2	0.011	0.044	.143**	.423**
通鼎光电	7	0.008	0.039	.487**	.743**
永鼎股份	15	0.008	.216**	.372**	.677**
南京熊猫	6	0.006	0.065	.201**	.427**
新海宜	4	0.005	.153**	-.278**	.402**
国腾电子	6	-0.008	.155**	-0.061	.323**
亿阳信通	27	-0.009	.162**	-.212**	.820**
*ST 华光	11	-0.015	0.088	.175**	.311**
东方通信	28	-0.016	.328**	.671**	.707**
华星创业	2	-0.018	0.105	0.014	.239**
星网锐捷	11	-0.026	.343**	.114*	.442**
闽福发 A	7	-0.035	.165**	.107*	.522**

表 4-8（续表）

股票	平均日股评数量	简单情感指数与收盘价	简单情感指数与成交量	股评数量与收盘价	股评数量与成交量
ST 太光	5	-0.043	0.029	.251**	.391**
世纪鼎利	17	-0.047	-0.028	.242**	.370**
辉煌科技	5	-0.047	.236**	-.152*	.739**
拓维信息	12	-0.051	.225**	-.354**	.849**
二六三	8	-0.052	0.026	.323**	.223**
深桑达 A	8	-0.057	-0.074	.371**	.560**
佳都新太	28	-0.066	.166**	.263**	.599**
上海普天	8	-0.071	.253**	.388**	.635**
汇源通信	23	-0.078	-0.001	.642**	.768**
海格通信	21	-0.078	-0.107	.629**	.641**
天音控股	34	-0.086	.221**	.405**	.403**
三维通信	4	-0.091	0.051	.314**	.574**

**. 在 0.01 水平（双侧）上显著相关。*. 在 0.05 水平（双侧）上显著相关。

4.3.2 股评变量因素方差分析

为证明股评变量（情感指数、股评数量）上的差异对股票变量的影响，本部分通过单因素方差分析来考察股票变量对股票表现的影响。

采用东方财富网通讯行业在 2011 年 12 月的数据，验证股评变量对股票表现的影响。利用简单情感指数、及股评数量数据作为股评变量来分析其对股票表现的影响。其中，每支股票股评的简单情感指数为 12 月所用简单情感指数的平均值，股评数量采用 12 月所有开盘日的平均股评数量。股票变现变量为 12 月的交易量、波动率、收益率、涨跌额度。

为控制各股票之间其他影响股票表现的变量，本文采用聚类方法，对各股票进行聚类，力求实验中个股票其他因素基本相同。基本面信息是影响股票表现的重要信息，本文通过对各股票 2011 年 12 月的会计信息（每股收益、应收账款周转率、资产报酬率、销售利润率、存货周转率、速动比率、流动比率，该数据来自国泰安数据库）利用 K 均值法进行聚类，并将类别最大的一组包含 50 支股票作为因素方差分析的数据。

按照平均每日股评数量将个股票分为三组，小于 4 的为第一组，包含 12 支股票，大于等于 4 小于 8 的为第二组，共有 18 支股票，大于等于 8 的为第三组，有 20 支股票。因素方差分析结果如表 4-9 所示。股评变量对股票市场波动率、交易量、收益率在 0.05 的水平下有显著影响，而在 0.1 的显著水平下对涨跌额的绝对值有显著影响，而对收益率的影响并不显著。

表 4-9 **ANOVA**

		平方和	df	均方	F	显著性
换手率	组间	.898	2	.449	4.833	.012
	组内	4.366	47	.093		
	总数	5.264	49			
交易量	组间	1.485E17	2	7.424E16	4.464	.017
	组内	7.816E17	47	1.663E16		
	总数	9.301E17	49			
波动率	组间	.003	2	.002	.809	.451
	组内	.088	47	.002		
	总数	.091	49			
收益率	组间	.076	2	.038	3.207	.049
	组内	.555	47	.012		
	总数	.631	49			
涨跌额	组间	11.546	2	5.773	3.078	.055
	组内	88.159	47	1.876		
	总数	99.705	49			

对股票评论从数量上对股票市场表现影响分析之后，发现其对股票表现比较显著，接下来本文将对股票评论所包含的情感信息对股票表现的影响进行分析。通过股评变量与股票表现相关分析，结果发现情感指数与股票收盘价并不都相关，在本部分实验中只选择股评情感指数与股票价格相关的支股票，剔除一支在上述聚类过程中排出的股票，剩余 19 支股票，根据 2011 年 12 月的股评平均每日的简单情感指数的大小，将股票分为 3 个组，第一组的三支股票情感指数为积极，第二组 10 支股票情感指数为中性，第三组 6 支股票情感指数为消

极，因素方差分析如表 4-10 所示。情感指数对股票波动率及收益率在 0.01 的水平下显著，对交易量、涨跌额、换手率没有显著地影响。

表 4-10 **ANOVA**

		平方和	df	均方	F	显著性
交易量	组间	2.415E16	2	1.208E16	.294	.749
	组内	6.565E17	16	4.103E16		
	总数	6.806E17	18			
波动率	组间	.016	2	.008	12.692	.000
	组内	.010	16	.001		
	总数	.027	18			
收益率	组间	.057	2	.028	6.262	.010
	组内	.073	16	.005		
	总数	.130	18			
涨跌额	组间	1.737	2	.869	.425	.661
	组内	32.689	16	2.043		
	总数	34.427	18			
换手率	组间	.001	2	.000	.089	.916
	组内	.083	16	.005		
	总数	.084	18			

4.4 大盘研究

4.4.1 股评变量与大盘表现相关分析

上证综合指数是基于上海证券交易所挂牌上市的所有股票进行计算而来，能够反映上海证券交易市场的总体趋势，是众多投资者参考的最重要指标之一。由于其影响程度重大，影响范围广泛，关注人数众多，股评数量巨大，对该指数的评论进行研究，了解其参与者的心理及情绪意义重大。

2010 年 11 月 8 日至 2012 年 4 月 13 日，共 348 个交易日，包括股票评论

23,2582 条股票评论，剔除非交易日的股票评论后剩余 20,0061 条评论，平均每天评论数为 573 条。经过无关评论清除、评论情感倾向分析得到无关评论 8,8879 条，相关评论 11,1047 条，其中看涨评论 5,2143 条，看跌评论 5,8904 条。

将股票评论变量与股票市场变量进行 Pearson 相关分析，分析结果如表 4-11 所示，收盘价与股评变量在 0.01 的显著水平下相关系数为 0.669，相关性较强，只与情感指数变量中的看涨指数在 0.05 的显著水平下相关，且相关系数不强，这与单只股票的结果差异较大。单只股票中收盘价与当天的看涨指数、简单情感指数相关，且相关性要强于大盘与情感指数的相关性。交易量及收益率均与看涨指数，简单情感指数及股评数量均在 0.01 的水平下显著相关。而波动率股评数量相关，这与单只股票是一致的。

表 4-11 相关分析

	收盘指数	交易量	波动率	收益率
看涨指数	-.115*	.174**	-.021	.213**
简单情感指数	.038	.229**	-.061	.213**
情感差异	.022	.081	.052	.029
股评数量	-.669**	.487**	-.211**	-.171**

**. 在 0.01 水平（双侧）上显著相关。 *. 在 0.05 水平（双侧）上显著相关。

4.4.2 股评变量与大盘及单只股票相互影响分析

本文进一步对股票评论与滞后股票表现进行相关分析，结果发现三个情感指数中只有看涨指数与滞后收盘指数显著相关，且为负相关。为研究看涨指数与上证指数收盘价的相关性，本文将看涨指数与领先 10 天、滞后 41 天的股票收盘见进行相关分析，结果发现看涨指数与滞后的收盘价均在 0.01 水平下显著负相关，与当日及领先一至八天收盘价均在 0.05 的显著水平下呈负相关。相关系数变化如图 4-6 所示。随滞后时间增加，相关性逐渐增强，在滞后 29 天时大最大值，相关系数为-0.289，对于领先看涨指数，在领先三天使达到最大，为-0.119。由此可以说明投资者对大盘的情绪主要受过去一段时间的大盘情况的影响，而其对大盘未来走向的影响相对较小，且显著性有所降低。

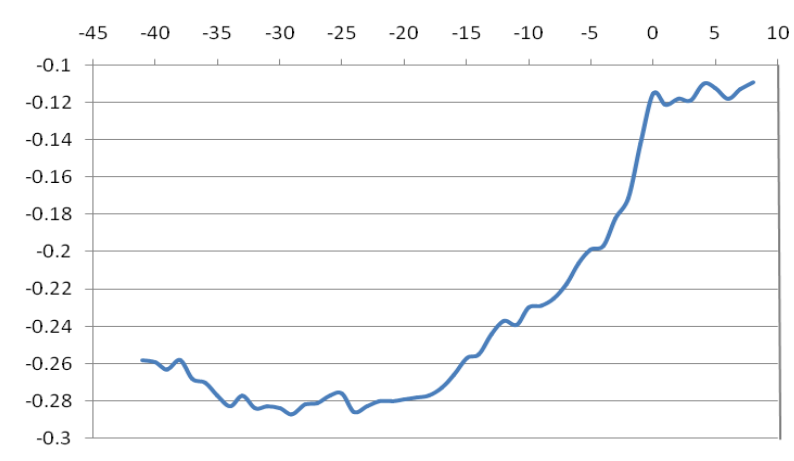

图 4-6 相关系数变化图

本文对单只股票，以中国联通为例，股票收盘价与股票评论情感指数进行相关分析，结果发现，两个情感指数（看涨指数、简单情感指数）与滞后收盘价均不显著相关，而与当日及领先收盘指数存在相关关系，而对于看涨指数除与当日收盘价在 0.01 的水平下存在正相关关系，在领先一至八天都在 0.05 的水平下显著相关，且随领先天数增加，相关系数逐渐减小。而简单情感指数与单日及领先情感指数均在 0.01 的显著水平下显著相关。相关系数变化如图 4-7 所示，由于股评数量在领先先第七天开始，也显著与股票收盘价相关，故也将其列入图中，图中所示的所有值均为在 0.01 或 0.05 的显著水平下显著相关的相关系数。简单情感指数与收盘价在当日相关性最强，随时间推移，相关性逐渐减弱，在领先 22 天时，不再显著相关。而股票数量在领先第七天开始才与收盘价相关，且相关系数均为负数，随领先天数增加相关性逐渐增强，在第 18 天时，相关性最强为-0.166，之后相关性开始下降，至领先 29 天以后不再显著相关。

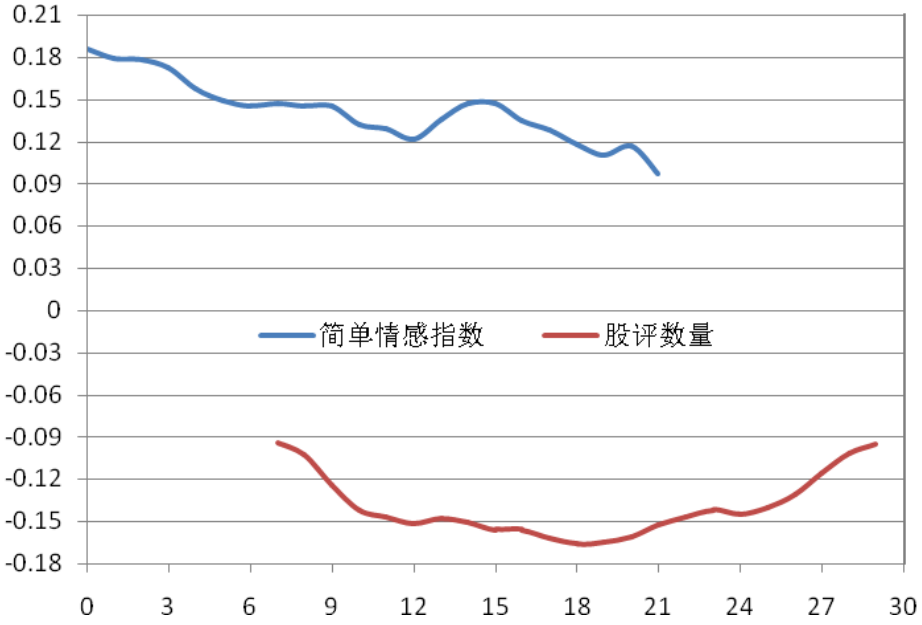

图 4-7 中国联通股评变量与领先收盘价相关分析

通过以上分析，股评变量与滞后大盘指数显著负相关，而与领先大盘指数相关性较弱，说明股评变量受大盘表现情况的负向影响，即当大盘表现越好时，投资者的情绪会越差，而当盘表现较差时，投资者会更乐观。

而对于单只股票，股评变量仅与当日及领先收盘价相关，与滞后收盘价不显著相关，说明股评变量能够影响未来收盘价，而不受过去收盘价的影响。简单情感指数对领先时间增加，收盘价的相关关系下降，而二者的相关系数为正，说明当投资者持积极情绪时，股票价格有上扬趋势，而投资者持消极情绪时，股价有下跌趋势。而股评变量会对股票价格产生消极影响，但影响并不会马上显现，而是在 6 至 7 天之后影响开始逐渐增大，并在 18 天时达到最大值，之后影响又逐渐减小。

综上论述，投资者的情绪能够受到大盘表现的影响，但对大盘影响较小，而不会受到单只股票的表现的影响，但能够影响单只股票的表现。

4.5 本章小结

本章首先对不同网站的股票论坛的参与者行为进行对比，发现不同网站的参与者行为基本相似。分析了股票价格的影响因素，并建立股票价格回归预测

模型，对股票价格的预测准确率为 98.55%，发现当日的机构评分、投资者情绪及前一日的公司新闻数量对股票的收盘价有显著的影响。通过信息增益及卡方特征提取方法对影响股票价格的可能因素进行分析，发现简单情感指数、看涨指数、滞后一天新闻数量是影响股票价格走向的重要因素。将股票评论变量与通讯行业股票进行相关分析、因素方差分析，发现简单情感指数对股票收益及波动率有相助影响，股评数量对收益率、换手率、涨跌额、交易量有显著影响。将股评变量与大盘及单只股票进行领先、滞后收盘价进行相关分析，发现大盘能够影响投资者情绪，而投资者情绪能够影响个股表现。

结　论

随着互联网技术发展及普及，网络成为重要的信息交流平台，众多的投资者利用网络获取信息，与其他投资者共同交流投资经验，并在网络上发表自己的观点，这给我们提供了一个了解投资者的途径。本文利用投资者在参与股票论坛过程中产生的文本信息对投资者的心理及行为进行研究。

本文通过文本分类及情感分析方法对股票评论进行分析，从非结构化的评论中提取股票市场相关信息，并进一步分析该信息的情感倾向。利用文本分类技术能够提取 84%的无关评论，并保留了 90%以上的股票市场有用评论，充分实现了股票市场无关评论剔除。情感分析过程中对比了语义方法、机器学习方法及 N-Gram 方法，结果发现利用 SVM 结合信息增益的机器学习最优的情感分析方法。

本文分别对单只股票、通讯行业及大盘的股票评论进行情感分析，并利用股评变量对股票市场进行研究。通过对单只股票价格影响因素分析，建立股票价格预测模型，发现滞后收盘价、简单情感指数、滞后一天新闻数量、机构评级是影响股票价格的重要因素，对股票价格预测准确率为 98.5%。进一步分析影响股票价格走向的因素，发现看涨指数、简单情感指数、滞后一天新闻数量对走向有预测作用。通过对通讯行业股票单因素发差分析，发现简单情感指数是影响股票收益率及换手率的重要因素。通过对比大盘、单只股票股评变量与领先、滞后收盘价进行对比发现，大盘走向是影响投资者情绪的变量，而投资者情绪能够影响单只股票的表现。

本文的主要研究成果及创新点归纳如下：

1．本文提出了领域无关信息的自动剔除方法，并成功剔除了股票市场无关信息。

2．应用语义分析方法、机器学习方法及 N-gram 方法对中文股票论坛进行情感分类，研究发现利用 SVM 结合信息增益的机器学习方法是三种方法中较好的。

3．通过对单只股票价格影响因素分析，建立了股票价格预测模型，能够比较准确地预测股票市场的价格。

本文利用了机器学习方法进行情感分析，需要大量已标注类别的训练数

据，且随时间推移，训练样本需要更新。本文提出的预测模型仅在单只股票两年的数据进行了验证，可扩展到多只股票、更长时间段证明其普适性.

参考文献

[1] 中国上市公司市值管理研究中心（CCMVM）. 2011 年度中国 A 股市值年度报告. 2012.

[2] Shleifer A. Inefficient markets: An introduction to behavioral finance[M]. Oxford University Press, USA，2000： 45-47.

[3] Shleifer A,Summers L H. The noise trader approach to finance[J]. The Journal of Economic Perspectives, 1990. 4(2): 19-33.

[4] Graham B,Dodd D. Securities Analysis[J]. PRINCIPLES AND TECHNIQUES (4th), 1934.

[5] JohnWiiliams. JohnWiiliams[M]. 1938： 35-42.

[6] Rhea R. 道氏理论[M]. 1932： 23-24.

[7] Kahneman D,Tversky A. Prospect theory: An analysis of decision under risk[J]. Econometrica: Journal of the Econometric Society, 1979: 263-291.

[8] Dreman D N. Contrarian investment strategy: The psychology of stock market success[M]. Vintage Books，1981： 51-58.

[9] Grinblatt M,Titman S,Wermers R. Momentum investment strategies, portfolio performance, and herding: A study of mutual fund behavior[J]. The American economic review, 1995: 1088-1105.

[10] Fisher K L,Statman M. A behavioral framework for time diversification[J]. Financial Analysts Journal, 1999: 88-97.

[11] Banz R W. The relationship between return and market value of common stocks[J]. Journal of Financial Economics, 1981. 9(1): 3-18.

[12] Fama E F,French K R,Prices U o C C f R i S. Dividends, Debt, Investment, and Earnings[M]. Center for Research in Security Prices, Graduate School of Business, University of Chicago，1997： 124-131.

[13] Baker M,Wurgler J. Investor sentiment in the stock market. 2007, National Bureau of Economic Research Cambridge, Mass., USA.

[14] Brown G W,Cliff M T. Investor Sentiment and Asset Valuation*[J]. The journal of Business, 2005. 78(2): 405-440.

[15] Kurov A. Investor sentiment and the stock market's reaction to monetary policy[J]. Journal of Banking & Finance, 2010. 34(1): 139-149.

[16] Brown G W,Cliff M T. Investor sentiment and the near-term stock market[J]. Journal of Empirical Finance, 2004. 11(1): 1-27.

[17] 高清辉. 论投资者情绪对股市的影响[J]. 经济纵横, 2005(4).

[18] Dailami M,Masson P. Measures of investor and consumer confidence and policy actions in the current crisis[M]. World Bank，2009： 23-27.

[19] Simon D P,Wiggins III R A. S&P futures returns and contrary sentiment indicators[J]. Journal of Futures Markets, 2001. 21(5): 447-462.

[20] Bandopadhyaya A,Jones A L. Measures Of Investor Sentiment: A Comparative Analysis Put-Call Ratio Vs. Volatility Index[J]. Journal of Business & Economics Research (JBER), 2011. 6(8).

[21] Lee W Y,Jiang C X,Indro D C. Stock market volatility, excess returns, and the role of investor sentiment[J]. Journal of Banking & Finance, 2002. 26(12): 2277-2299.

[22] Neal R,Wheatley S M. Do measures of investor sentiment predict returns?[J]. Journal of Financial and Quantitative Analysis, 1998. 33(4).

[23] Baker M,Wurgler J. Investor Sentiment and the Cross‐Section of Stock Returns[J]. The Journal of Finance, 2006. 61(4): 1645-1680.

[24] 李潇潇,杨春鹏. 投资者情绪和 A 股溢价的关系研究[J]. 统计与决策, 2009(20): 124-126.

[25] 池丽旭,庄新田. 投资者情绪与股票收益波动溢出效应[J]. 系统管理学报, 2009(04): 367-372.

[26] Wysocki P. Cheap talk on the web: The determinants of postings on stock message boards[J]. Working Paper, Univercity of Michigan, 1999.

[27] Wysocki P D. Private Information, Earnings Announcements and Trading Volume, or Stock Chat on the Internet. A public Debate about Private Information, Working Paper, University of Michigan Business School[J]. Ann Arbor, 1999. 1001: 48109-1234.

[28] Tumarkin R,Whitelaw R F. News or noise? Internet postings and stock prices[J]. Financial Analysts Journal, 2001: 41-51.

[29] Antweiler W,Frank M. Internet stock message boards and stock returns[J]. University of British Columbia Working Paper, 2002.

[30] Sabherwal S,Sarkar S K,Zhang Y. Online talk: does it matter?[J]. Managerial Finance, 2008. 34(6): 423-436.

[31] Das S,Martínez‐Jerez A,Tufano P. eInformation: A clinical study of investor discussion and sentiment[J]. Financial Management, 2005. **34**(3): 103-137.

[32] 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.

[33] Sehgal V,Song C. SOPS: stock prediction using web sentiment: 2007, IEEE: 21-26.

[34] Zimbra D,Fu T,Li X. Assessing public opinions through Web 2.0: a case study on Wal-Mart. Proceedings of the 30th International Conference on Information Systems[C]. 2009.

[35] Koski J L,Rice E M,Tarhouni A. Noise trading and volatility: Evidence from day trading and message boards[J]. University of Washington, 2004.

[36] Oh C,Sheng O. Investigating Predictive Power of Stock Micro Blog Sentiment in Forecasting Future Stock Price Directional Movement. International Conference on Information Systems[C]. Shanghai: 2011.

[37] Fuller R J. Behavioral finance and the sources of alpha[J]. Journal of Pension Plan Investing, 1998. 2(3): 291-293.

[38] 李心丹. 行为金融理论: 研究体系及展望[J]. 金融研究, 2005(001): 175-190.

[39] Hong H,Stein J. A unified theory of underreaction, momentum trading, and overreaction in asset markets[J]. The Journal of Finance, 1999. **54**(6): 2143-2184.

[40] Barberis N,Shleifer A,Vishny R. A model of investor sentiment1[J]. Journal of Financial Economics, 1998. 49(3): 307-343.

[41] Daniel K,Hirshleifer D,Subrahmanyam A. Investor psychology and security market under‐and overreactions[J]. The Journal of Finance, 1998. **53**(6): 1839-1885.

[42] Black F. Noise[J]. Journal of Finance, 1986. 41(3): 529-43.

[43] 王冀宁,吴启宏,李心丹. 中国股票市场机构与散户的均衡策略及实证研究[J]. 当代财经, 2004(08): 30-34.

[44] Pang B,Lee L,Vaithyanathan S. Thumbs up?: sentiment classification using machine learning techniques: 2002, Association for Computational Linguistics: 79-86.

[45] Turney P D. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. Annual Meeting on Association for Computational Linguistics[C]. 2002, Association for Computational Linguistics: 417-424.

[46] 刘开瑛. 中文文本自动分词和标注[M]. 商务印书馆，2000： 156-159.

[47] 宋枫溪,郑如冰,王积忠. 自动文本分类中两种文本表示方式的比较[J]. 计算机工程, 2004(18): 124-126.

[48] Rosso P,Ferretti E,Jiménez D, et al. Text categorization and information retrieval using wordnet senses: 2004.

[49] 代六玲,黄河燕,陈肇雄. 中文文本分类中特征抽取方法的比较研究[J]. 中文信息学报, 2004(01): 26-32.

[50] Woon D E,Dunning T H. Gaussian basis sets for use in correlated molecular calculations. III. The atoms aluminum through argon[J]. The Journal of chemical physics, 1993. 98(2): 1358.

[51] 何晓群. 多元统计分析[M]. 中国人民大学出版社，2004： 117-125.

[52] 石志伟,吴功宜. 改善朴素贝叶斯在文本分类中的稳定性. NCIRCS2004 第一届全国信息检索与内容安全学术会议[C]. 中国上海: 2004: 137-146.

[53] 奉国和. 文本分类性能评价研究[J]. 情报杂志, 2011(08): 66-70.

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

[55] Antweiler W,Frank M Z. Is all that talk just noise? The information content of Internet stock message boards[J]. Journal of Finance, 2004. 59(3).

[56] 李子奈,潘文卿. 计量经济学 2 ed[M]. 高等教育出版社，2005： 56-58.

[57] Shmueli G. Predictive analytics in information systems research. 2010, University of Maryland.

攻读硕士学位期间发表的论文

[1] 李玉梅,闫相斌,胡洋. 在线股评对股票市场的影响分析. 统筹优选与经济转型——第十三届中国管理科学学术年会[C]. 中国浙江杭州: 2011: 386-390.

哈尔滨工业大学学位论文原创性声明及使用授权说明

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《基于互联网的股票市场趋势预测》，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。据本人所知，论文中除已注明部分外不包含他人已发表或撰写过的研究成果。对本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。本声明的法律结果将完全由本人承担。

作者签名：李玉梅　　日期：2012年7月8日

学位论文使用授权说明

本人完全了解哈尔滨工业大学关于保存、使用学位论文的规定，即：

（1）已获学位的研究生必须按学校规定提交学位论文；（2）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；（3）为教学和科研目的，学校可以将学位论文作为资料在图书馆及校园网上提供目录检索与阅览服务；（4）根据相关要求，向国家图书馆报送学位论文。

保密论文在解密后遵守此规定。

本人保证遵守上述规定。

作者签名：李玉梅　　日期：2012年7月8日

导师签名：[illegible]　　日期：2012年7月8日

致谢

非常感谢我的导师闫相斌教授，从文献查阅、论文选题、研究方法确定、研究工作的进行都是在导师的悉心指导下完成的。在两年的导师严谨的治学态度、渊博的知识、宽以待人的优秀品德给我留下了深刻的印象，让我感受到教书育人的学者风范，是我在今后学习、工作和生活中的楷模。在这里谨向我的导师致以崇高的敬意和深深的感谢。

感谢管理科学与工程所有老师给我提出的宝贵意见和建议，论文的成稿与他们的指导是分不开的。

感谢和我朝夕相处的同学们，感谢室友们，和他们在一起，我度过了充实快乐的两年硕士生活。

感谢我的父母对我的理解和支持，是他们的关心和鼓励，使我顺利完成学业。

最后预先向在百忙之中抽出宝贵时间审阅论文和参加我的答辩的各位老师表示衷心的感谢！