学校代码：10272

学　　号：

上海财经大学

应用统计硕士学位论文

基于情感分析探究网络财经新闻对股票价格走势的影响

院（系所）　统计与管理学院

专　　业　　应用统计硕士

姓　　名

指导教师

完成日期　2016 年 3 月 28 日

学位论文原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不含任何其他个人或集体已经发表或撰写过的作品成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本人完全意识到本声明的法律结果由本人承担。

学位论文作者签名：　　日期：　　年　　月　　日

学位论文版权使用授权书

本人完全了解上海财经大学关于收集、保存、使用学位论文的规定，即：按照有关要求提交学位论文的印刷本和电子版本；上海财经大学图书馆有权保存学位论文的印刷本和电子版，并提供目录检索与阅览服务；可以采用影印、缩印、数字化或其它复制手段保存论文；在不以赢利为目的的前提下，可以公布论文的部分或全部内容。（保密论文在解密后遵守此规定）

论文作者签名：　　导师签名：

日期：　　年　　月　　日　　日期：　　年　　月　　日

摘要

科技的发展不仅仅让人们的生活愈来愈便捷，也让我们接触信息的渠道不断增多。互联网发展到今天，已经深入到各个领域，人们的工作与生活方式也不可避免地受到冲击，过去的报纸等信息传播渠道逐渐被互联网所代替。随着互联网的发展，越来越多的文本信息出现在网络中,这些信息不仅对商业，对经济，甚至对政治和文化有很深远的影响。中国证券市场经过了二十多年的发展，股民人数暴增。对于广大股民而言，互联网财经新闻中所包含的很多有价值的数据与文本信息都与他们的利益息息相关。他们所关注的文本信息或者是关于公司的某一重大举措，或者是某个专家关于公司的评论，亦或是公司的某个丑闻。这些内容或多或少都会有一定的情感倾向，影响股民对于公司股票未来走势的预期，进一步影响股票价格走势。并且已有研究表明投资者的投资行为会受到互联网财经新闻信息的影响，因此研究互联网财经新闻与股票价格波动趋势之间的关系是一项很有意义的研究。

过去在这方面的研究大多使用如支持向量机的方法处理文本数据，而本文将创新性的利用情感分析方法来研究从互联网上得到的财经新闻文本信息。本文旨在研究以中国银行等五家上市银行为代表的目标企业在一段时间内出现在以新浪财经为代表的目标新闻源上的所有个股新闻所能传达出的情感倾向，以及这些情感倾向会对股票市场价格的波动产生怎样的影响。该分析主要包括两方面：首先是利用文本挖掘技术以及情感分析对所得到的关于中国银行(601988)的个股新闻标题进行分词以及情感分析，从而得到每条新闻标题的情感得分并进行一定的处理；其次是应用上一步骤得到的每只股票每个交易日的情感得分、相应的隔夜 SHIBOR、即期汇率、上证指数、银行指数以及对应股票市场价格变化的波动情况，分别进行 Logistic 回归分析以及自相关 Logistic 回归分析并利用最终得到的模型用于预测。得到的结论是新浪财经上发布的个股新闻对应的情感得分对该日股票价格是否产生涨幅有显著影响。其中，五家银行对应的模型中情感得分的参数估计值都大于 0，这说明新闻所表达的情感倾向越积极，对应的情感得分越高，则相应的该日股票价格发生涨幅的概率越大，这一结论与现实上的想法是一致的。另外通过模型的预测检验结果可以发现，模型得到的结

果比较保守，消极的、不利于公司的新闻会比较容易直接影响股票价格的波动，而当面对积极的新闻时，投资者更多时候可能会持一种暂时观望的态度，不会立即采取行动。

本文的内容结构安排如下：第一章介绍论文的研究背景、意义以及研究方法；第二章将对本文设计的相关文本处理技术进行介绍；第三部分重点介绍本文使用的 Logistic 回归的相关理论知识；第四章是本文的数据准备以及实证研究部分，介绍研究的步骤以及结果；第五章将针对研究结果总结相应的结论并提出相关建议，同时提出本文研究上的不足之处及未来的改进方法。

关键词：网络财经新闻，文本挖掘，情感分析，Logistic 回归

Abstract

The development of technology does not only make human's life more convenient, but also enrich the way that people receive information. The internet has developed so deeply in any kind of field that the patterns of people's work and life have been inevitably struck. As the developing of the Internet, more and more text information occurs on the Internet, which has a far-reaching impact on commerce, economy even political and culture. Until now, Internet-based news often comes at the right time and it is very easy to access, so it becomes more and more popular. Especially, Internet-based financial news tells something important about the operation or financial condition of a company, which is very valuable for Chinese investors. Some research has already shown that the investment behavior of investors is influenced by Internet-based financial news information. Thus, it is very meaningful to study the relationship between Internet-based financial news and the fluctuation tendency of stock price.

Instead of relying on support vector machine, this paper will choose the method of sentiment analysis to deal with text data. This paper focuses on the study of sentiment tendency conveyed by the target stock news appearing on the Internet represented by SINA Finance during the giving date, and how this sentiment tendency affects the fluctuation of stock price. In this study, the sentiment analysis is used to quantify the stock news to be sentiment score. Then it is the logistic and autocorrelation logistic regression that we apply to model the fluctuation of stock price and independent variables such as sentiment score, SHIBOR, exchange rate SSEC and KBW. The conclusion is that the sentiment score is positively correlated with the fluctuation of stock price in the model, which means that the higher the sentiment score, the greater the probability of the increasing of stock price. This conclusion is in keeping with intuitive sense. On the other hand, the consequence of predicting reveals that the model is conservative and the negative stock news is more likely to directly affect the fluctuation of stock price than positive news.

Abstract

The whole paper is divided into six parts. The first part introduces the background of the research content related with this paper. The second part describes the theoretical basis and technique of text mining used in this paper in detail. The third part introduces the theoretical knowledge and method of estimation about logistic regression model. The fourth part describes the data preparations and the key part in this research. The last part introduces the summary, advice of this thesis, the deficiency and the improved method of this thesis.

Key Words: Internet-based financial news; Text mining; Sentiment analysis; Logistic regression

目录

第一章　引言

第一节　研究背景及意义

一、研究背景

科技的发展不仅仅让人们的生活愈来愈便捷，也让我们接触信息的渠道不断增多。二十世纪末，互联网媒体由于其传播速度快、容量大、自由度高和形式活泼新鲜等优点得到迅速发展与大规模普及。互联网作为中国经济社会发展的一个基础设施，是经济创新的重要要素。互联网发展到今天，已经深入到各个领域，人们的工作与生活方式也不可避免地受到冲击，过去的报纸等信息传播渠道逐渐被互联网所代替。而互联网新闻中涵盖了大量的信息，尤其是财经部分的新闻报道，不仅包含上市公司的财务数据，同时包含了很多关于上市公司相关的文本信息。已有文献显示，互联网财经新闻会对股票市场的价格波动产生一定的影响。

随着中国证券市场 20 多年的不断发展，活跃于中国股市中的投资者数量剧增。对于广大股民而言，互联网财经新闻中所包含的很多有价值的数据与文本信息都与他们的利益息息相关。股民所关注的文本信息或者是关于公司的某一重大举措，或者是某个专家关于公司的评论，亦或者是公司的某个丑闻。这些内容或多或少都会有一定的情感倾向，影响股民对于公司股票未来走势的预期，进一步影响股票价格走势。

数据挖掘发展到如今已经比较完善，越来越多的企业、政府部门甚至社会机构逐渐明白数据的价值与使用方法，也乐意花费大量资金在数据挖掘的研究中。而现实生活中，除了数据以外还有大量的文本信息并没有被很好的挖掘与利用。数据挖掘技术中外通用，但是文本挖掘相关技术则外国明显先进于中文。

二、研究意义

作为信息时代公众获取财经信息的主要渠道，探究互联网财经新闻与股市波

动的具体关系是非常有意义的。本文欲利用互联网财经新闻，不论是正面的还是负面的，对股票价格的走势进行一定的预测。若能将这些信息用于资本市场投资操作等行为上都是很有价值的，也希望能得到有意义的结论以此为投资者与管理者提供一些决策建议。

文本挖掘是新时期人们对数据的更加深入的需求，研究的原材料是各种文本格式的文字、图片等，通过这些来分析相似、关键性和内部蕴涵的逻辑结构等等。随着互联网的发展，越来越多的文本信息出现在网络中,这些信息不仅对商业，对经济，而且对政治和文化有很深远的影响。但是这一部分的中文文本信息相对数据来说比较难以研究，但其价值相比数据可以说是不相上下的，因此具有很大的研究意义。

第二节　文献综述

目前为止，有大量的相关研究表明新闻对于股价波动是有影响的。因为新闻中通常涵盖大量与上市公司相关的信息，无论对于投资者还是分析者，分析这些信息对股票价格波动的影响都是很有意义的。

一、国外研究综述

对于网络财经新闻与股票价格波动的研究国外很早就已经开始了，从二十世纪七十年代开始，有关新闻对于资产交易价格波动影响的研究文献就不断大量涌现。

Niederhoffer (1971)[22]搜集了 1950 年至 1966 年发表于《纽约时报》上的 432 篇重大新闻，并利用统计的方法研究分析新闻发布后股票市场指数的涨跌幅情况。作者发现股票市场在新闻发布的当天反应最为强烈，并且第一天市场指数的变化往往会影响第二天指数的变化趋势。之后他进一步根据新闻的标题内容人工将其分为 20 类，以此研究不同类别的新闻对市场波动的影响情况。

Wiithrich(1998)[29]开发了一个可以对每天的股票市场指数上涨、下跌还是稳定进行预测的股票指数预测系统，该系统还可预测当天收盘时的股票指数。作者的研究步骤主要是，首先分析总结新闻中的关键字以及每个关键字出现后股票市场指数的走势；然后依据每个关键字在不同走势下出现的次数来计算权重，

根据统计结果来分析关键字出现次数与股票市场指数走势之间的规则函数，并且利用这些规则来预测其他新闻报道后的股票指数走势。

Tetlock(2008)[28]试图研究除了直观的会计变量和分析者的分析结果之外的财经新闻对于企业股票价格变动的影响。作者在研究中使用的方法正是本文尝试采用的情感分析，不同的在于英文在这方面的研究相较中文更加简单与先进，其中包括分词的简单以及词库的完整性。作者在文章中利用了Harvad − IV − 4词库，根据研究新闻中积极和消极词汇的词频，将新闻分为正面与负面新闻，以此来研究新闻是否可以预测单个企业会计收益和股票价格。其研究的第一个发现是对于分析者的预测和历史会计数据来说负面新闻相通常更可能预示着企业的低收益，第二个发现是负面新闻的产生会对企业的股票价格有一个延迟期为一天的小的影响。

Maragoudakis(2015)[21]利用 MCMC 贝叶斯推断方法研究了金融新闻与社会媒体舆论对股票市场的影响。由于人工获得一个公司大量新闻太过繁琐，本文只用虚拟交易结果验证了该方法。因此并不代表现实生活中的金融新闻会对股票市场产生影响。

二、国内研究综述

相对于国外学者在互联网新闻与股票价格关系上的研究，国内学者相对比较滞后也比较少。但是随着我国网络的飞速普及，股票市场的发展以及中文文本挖掘上不断深入的研究，这些因素促使国内越来越多的学者针对我国网络财经新闻对于企业股票市场影响的研究。

杨继东(2007)[10]在文章中对媒体影响资产价格的文献进行了梳理综述，在理论上分析讨论了媒体对投资者行为的影响，这一影响又是如何对资产价格产生影响的。其研究表明媒体可以影响投资者的情绪，而投资者的情绪通常会对其行为造成一定的影响，投资者的行为就包括参与度、投资选择和对收益的预期等这些会影响资产价格的因素。但是此研究仅限于理论层面，并没有通过实验或者实证研究证明其合理性。

赵伟和梁循(2009)[16]在文章中以网络金融信息的条数来代表信息流的强度，以此来研究网络金融信息量与股票收益率波动的关系。文章表明股票收益率与网络金融信息量成正比，即网络金融量的增加会带动股票收益率的增加，这也

就是说网络金融信息量会对股票收益率产生显著影响。

饶育蕾和王攀(2010)[12]在网络上搜集了两百多只股票的相关新闻，然后以发行价格与累计异常收益率作为被解释变量，以股票的新闻数量代表媒体关注度作为解释变量以及多个相关控制变量分别进行回归分析。文章表明股票的新闻数量会对股票产生短期的正向影响、长期的负向影响以及股票的发行价格有积极影响。

这三篇关于媒体新闻对于股票或是资产价格影响的文献只是从理论或是通过新闻数量研究了两者的关系，都完全没有考虑到了新闻本身所涵盖的文本信息。

国内利用文本挖掘技术来研究财经新闻的学者比较早的是陈华和梁循(2006)[1]，他们利用了统计词频的简单方法来统计个股新闻中出现其他股票的频数，并以此来研究分析某篇个股新闻对其他个股的影响，这算是比较简单的使用文本挖掘的方法。之后，赵丽丽(2012)[15]基于计算机科学与经济学，并利用计量经济学中多元回归与文本挖掘技术中的支持向量模型，不仅将文本信息量化为影响股市的因子而且得到其对股市的影响模型。其文章表明沪深两市的新闻都会对上市公司的股票产生影响，且对沪市股票的影响弱于深市的影响。

第三节　研究内容及特色

本论文将采用理论研究与实证分析相结合的方法：在理论研究中，本文将对文本挖掘、情感分析和 Logistic 回归分析等相关理论做出归纳和分析；在实证分析中，本文的重点在于以中国银行 (601988)等五家银行业上市公司为分析主体，研究一段时间内在目标新闻源上分别关于各个上市公司的新闻报道对于股票价格走势的影响，并通过 Logistic 回归分析建立预测模型。

本文主要内容包括：第二章，介绍相关的理论知识，包括文本挖掘、情感分析以及网页爬虫的一些知识；第三章，着重于介绍 Logistic 回归分析的相关理论基础和估计方法，并探究了 Logistic 回归分析在互联网财经新闻对股票的影响研究上的适用性；第四章，根据选择的上市公司以及财经网站，利用爬虫技术在研究的新闻源上获得实际文本数据进行实证研究，其中利用中文文本挖掘以及情感分析对这些新闻标题进行分析研究，挖掘出其中的有用信息即所体现出的情感倾向；然后通过 Wind 资讯金融终端将研究时间窗口内该公司的股票价格

等信息提取出，运用 Logistic 回归进行实证研究；第五章，总体回顾本文的主要结论，并在此基础上对投资者、企业管理者和政府提出了一些针对性建议，同时提出了本文的不足并以此得到改进方向。

在文献综述中可以看到，国内学者对互联网财经新闻对股票的影响研究比较晚，在文本挖掘这一块中文的发展也还远没有到完善的地步，同时运用情感分析在这方面进行的研究更少。因此本文的主要特色是将情感分析方法以及自相关 Logistic 回归分析引入到国内互联网财经新闻中进行研究。

第二章　文本挖掘相关技术

第一节　文本挖掘

一、文本挖掘的定义

在现实生活中，可获取的很多信息是以文本而非数据形式存储在文本数据库中的，如新闻文档、研究论文、书籍、电子邮件和网页页面等。而现在由于网络的飞速发展，电子形式的文本信息不断膨胀，文本挖掘也成为信息领域的研究热点。互联网时代，网络中充斥了海量的信息，企业需要对它们进行合理及有效的处理与利用，从而帮助企业在业务以及内部管理及发展等方面做出及时、正确的判断，在此基础上采取明智的行动，从而在竞争中占据主动权。众所周知的“啤酒和尿布”的故事正是利用了数据挖掘技术为企业带来很大收益的经典案例。在现实生活中，每天各个领域都在产生大量的数据与文本，数据的利用已经达到了比较细致的地步，而文本由于其获取以及分析的困难则较少的能够被挖掘出价值。

文本挖掘是数据挖掘的一种，但研究的对象不是数据而是无结构或者半结构化自由开放的文本。大致的意思就是对文本信息进行加工处理，对提取的信息进行统计处理的一项技术，将数据挖掘的成果用于分析以自然语言描述的文本。广义上说，只要对文本信息进行探索的分析均可以归类为文本挖掘。通常来说文本挖掘过程需要不断的尝试去获取最优模型，而不是一次就能完成，所以该过程需要不断调整。文本挖掘旨在通过识别和检索令人感兴趣的模式，进而从其数据源中抽取有用的信息。文本挖掘的数据源是文本集合，令人感兴趣的模式不是从形式化的数据库记录里发现，而是从非结构化的数据中发现。

二、文本挖掘的过程与中文分词

（一）文本挖掘的过程

（1）文本预处理：选取目标文本并将其转化为文本挖掘工具可以处理的中

间形式；

（2）文本挖掘：在完成文本预处理之后，利用机器学习、自然语言处理、数据挖掘以及模式识别等方法提取面向特定应用目标的知识或模式；

（3）模式评价与表示：这为最后一个环节，是利用已定义好的评估指标对获取的知识或模式进行评价。

（二）中文分词

与英文不同单词之间以空格作为自然分界不同，中文的基本书写单位是字，而词语之间则没有明显的分界，因此，中文的分词更加麻烦。中文分词作为文本挖掘的基础，对于目标文本成功的分词将会直接影响文本所表达出来的信息。

本文使用的用于分词的编程语言是 R，所利用到的软件包是Rwordseg。该软件包基于中科院的ictclas中文分词算法，并引用了孙健开发的ansj中文分词工具，可以用于人名、地名和组织机构的识别，以及多级词性标注和关键词等的提取。在此基础上，还支持用户自定义词典以及手动载入专业词典，这些都可以大大增强此软件包的分词效果。

第二节　情感分析

一、情感分析的定义

人与人的交流大多建立在语言的基础之上，语言的作用不仅仅是一些事实的传递，更多的时候是通过文本传递出不同的情感。情感分析是指用自然语言处理、文本挖掘以及计算机语言学等方法来识别和提取原材料中的主管信息，目的在于找出说话者/作者通过文本内容所表达的两极观点的态度。通常来说，情感分析分为两种，一类是主观性：主观、中性和客观；另一类是情感倾向：积极、中性与消极。以“这个手机很实用”为例，此句子的主要要表达的意思为“手机实用”，对于消费者来说，这个句子会带来关于这个手机积极的影响，从而促使他们产生购买这款手机的行为。

本文尝试利用的情感分析属于情感倾向研究，其基本任务是对目标文档或者句子等的情感极性进行分类，即文档或句子的情感倾向是积极、中性还是消极，并给予相应不同的情感得分。主要研究一段关于某只股票的新闻文本会对

股民产生一定的情感倾向影响，进而影响股民对于该股票的持有或者抛售的行为，从而带来股票价格趋势的变动。

语言交流是民众获得关于企业基本价值潜在的重要信息来源，因为很少有股票市场投资者会直接观察企业的生产活动，他们大多数的信息都是间接获得的。其中三个主要的来源为：分析师的预测、可量化的公开披露的会计变量以及关于企业当前和未来盈利的语言描述。如果分析师或者会计变量所提供的关于企业基本价值的测度是不完整或者有偏差的，那么语言变量可能会为企业未来的收益和回报增加解释力。

二、情感分析的研究方法

情感分析的研究方法主要分为两类：一种是基于情感词典的方法；一种是基于机器学习的方法。其中基于情感词典的方法需要用到标注好的情感词典。然而这类词典中英文的较多，中文的比较少。本文会使用到的词典包括：（1）台湾大学研发的中文情感极性词典 NTUSD；（2）知网发布的“情感分析用词语集”；（3）搜狗拼音输入法中提供的关于财经、金融、股票、品牌等方面的词库。其中前两个词库主要用于情感分类，而第三个词库则用于分词。

具体来说，针对某条财经新闻标题如“金融股表现低迷，西部证券领跌”。首先需要进行分词，所谓分词即为将一句话中的主谓宾表副词等分开，针对这个新闻标题，最理想的分词结果应该为“金融股”、“表现”、“低迷”、“西部证券”和“领跌”；然后分析这些词汇中带有情感倾向的是哪些词，很明显其中会带给读者一定情感倾向的词汇为“低迷”和“领跌”；接着判断情感倾向是积极、中性还是消极，可以感受到这两个词汇所表达出来的情感倾向都是消极的（注：如果一句中所有词汇都不表达积极或者消极的倾向，则认为此句子的情感倾向为中性）；最后对这个句子给定相应的情感得分，本句中总共有五个词汇，其中有两个会表达出消极情感的词汇，每个消极词汇给予-1分，因此本句的情感词汇得分为-2，另外考虑到其他不带有情感倾向词汇的影响，情感得分为$-2/5 = -0.4$.

本文并不认为语言的数量测度包括或是主导传统的企业基础价值的会计测量，而旨在探索企业新闻中消极词的分割是否会提高我们对企业现金流的了解以及股票的市场价格是否有效体现了这些文本信息。因为消极词统计数是定性

信息测度的扰动项，所以回归模型中的系数应该与零有一定的偏差，低估了定性信息的真正重要性。

第三节　网页爬虫及新闻内容提取

本文尝试针对多家股票价格波动以及新闻信息量较为完整的上市公司进行分析，挖掘出对应公司在研究时间内发生的财经新闻所传达的情感倾向，将其量化之后建立模型分析对该公司股票价格的具体影响并用于预测。因此本论文需要使用专门的网页爬虫，选取新浪财经作为新闻源，自动解析并抓取需要研究的上市公司在目标时间段内的新闻文本。

所谓的网页爬虫就是根据网页的 URL 完成页面目标文字数据等抓取的过程。而不同网站对应新闻网页的 URL 都有一定的规律。本文利用此规律，批量抓取目标新闻源上关于目标企业的财经新闻网页。由于本文的研究对象是新闻标题文本，因此需要对抓取的网页提取出研究时间段内的所有个股新闻标题。

本研究所用的用于爬虫的编程语言是 R，主要使用的软件包是RCurl。利用getURL函数将给定的 URL 上的网页内容抓取下来，抓取的时候需要分析网页源代码上的内容（见图 2.1 与图 2.2）

```
<!doctype html>
<html>
<head>

<title>中国银行(601988)行业资讯_新浪财经_新浪网</title>
<meta name="Keywords" content="中国银行行业资讯,601988行业资讯,新浪财经中国银行(601988)行
业资讯" />
<meta name="Description" content="新浪财经中国银行(601988)行情中心,为您提供中国银行
(601988)行业资讯信息数据查询." />
<meta http-equiv="Content-Type" content="text/html; charset=gb2312" />
```

图2.1 新浪财经源代码中的编码

图2.1上主要的内容是该网页的编码，主要看 charset 所显示的内容，从上图可以看到新浪财经的编码是 gb2312，因此每次抓取网页的时候都需要查看相应的源代码

```
    <td width="60"  valign="top">  <font style=" font-weight:bold">行业资讯:
</font></td></tr>
        <tr>
        <td>
        <div class="datelist"><ul>
                            2016-02-26 10:03  <a
target='_blank'
href='http://finance.sina.com.cn/stock/hkstock/ggscyd/20160226/100324347252.shtml'>冠君产
业全年每基金单位分派0.1985元</a> <br>    2016-02-
26 09:58  <a target='_blank'
```

图2.2 源代码中关于个股新闻的Path

抓取下目标网页之后，需要筛选出本文研究的个股新闻，而提取出相应部分信息则需要用XPath来解读。通过分析图 2.2 上显示的源代码，该网页个股新闻对应的XPath为//div[@class = 'datelist']/ul/a。得到对应的XPath后，即可利用 R 通过编写程序得到一定时间内的全部个股新闻标题。

第三章　Logistic 回归的相关理论知识

在回归模型中，如果解释变量为分类变量，可以通过引入虚拟变量进行处理。但是在社科研究中，经常会遇到因变量为分类变量的问题。例如，研究者关心的问题是哪些因素（性别，体重等）导致了人群中有些人患某种病而有些人不患某种病等。这类问题，实质上是一个回归问题，因变量就是上述提到的这些分类型变量，解释变量是与之相关的一些因素。但是，这样的问题却不能直接用线性回归分析方法解决，其根本原因在于因变量是分类变量。

首先，线性回归模型中“误差服从正态分布”的假设根本无法得到满足；其次，预测值通常会超出符合逻辑的可能范围(0~1)。Logistic 回归是针对定性变量进行的回归分析，可用于处理定性因变量的统计分析方法还有 Probit 分析、对数线性模型和判别分析(Discriminant analysis)等。而 Logistic 回归分析在社会科学中应用的最多，根据因变量取值类别的不同，该回归分析可以分为 Binary Logistic 回归分析（因变量只取 0 和 1 两个值）和 Multinomial Logistic 回归分析（因变量可以取多个值）。

第一节　模型介绍

首先考虑因变量的取值只有 0-1 两分类的情况，具有 p 个独立变量的向量 $X' = \{X_1, \cdots X_p\}$，设条件概率$Pr(Y = 1|x)$为根据观测量相对于某事件发生的概率。则 Logistic 回归模型可表示为

$$\pi(x) \coloneqq E(Y = 1|X = x) = Pr(Y = 1|X = x) \tag{3.1}$$

$$\text{logit}\big(\pi(x)\big) = \log\left(\frac{\pi(x)}{1 - \pi(x)}\right) = x'\beta = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \tag{3.2}$$

$$\pi(x) = \frac{e^{x'\beta}}{1 + e^{x'\beta}} = \frac{1}{1 + e^{-x'\beta}} \tag{3.3}$$

第二节　模型估计

设有 n 个观测样本，观测值分别为$y_1, y_2, \cdots, y_n$,设$p_i = P(y_i = 1|X = x_i)$为给定条件$X = x_i$下得到$y_i = 1$的概率。在同样条件下得到$y_i = 0$的条件概率为$P(y_i = 1|X = x_i) = 1 - p_i$。因此，可以得到一个观测值的概率为：

$$P(y_i) = p_i^{y_i}(1 - p_i)^{(1-y_i)} \tag{3.4}$$

因为各个观测独立，所以它们的联合分布可以变为各边际分布的乘积，即

$$L(\beta) = \prod_{i=1}^{n} \pi(x_i)^{y_i}[1 - \pi(x_i)]^{(1-y_i)} \tag{3.5}$$

(3.5)式被称为n个观测的似然函数，主要的目标是能够求出使得这一似然函数值最大的参数估计。因此，最大似然估计的关键在于求出参数$\beta_0, \beta_1, \cdots, \beta_p$的估计值，使得(3.7)式取得最大值。

对上述似然函数取对数得到

$$l(\beta) = \ln[L(\beta)] = \sum_{i=1}^{n} \{y_i \ln[\pi(x_i)] + (1 - y_i) \ln[1 - \pi(x_i)]\} \tag{3.6}$$

(3.6)式被称为对数似然函数，由于对数函数的单调递增性，最大似然估计即为了估计能使(3.6)式取得最大的参数$\beta_0, \beta_1, \cdots, \beta_p$的值。

对得到的对数似然函数求导，得到 p+1 个似然方程：

$$\sum_{i=1}^{n} [y_i - \pi(x_i)] = \sum_{i=1}^{n} \left[y_i - \frac{e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p}} \right] = 0 \tag{3.7}$$

$$\sum_{i=1}^{n} x_{ij}[y_i - \pi(x_i)] = \sum_{i=1}^{n} x_{ij} \left[y_i - \frac{e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p}} \right] = 0 , j = 1,2, \cdots, p \tag{3.8}$$

(3.7)与(3.8)式共同称为似然方程，为了解出上述非线性方程需要利用牛顿−拉夫逊(Newton − Raphson)方法进行迭代求解。

对l(β)求二阶偏导数，即得到Hessian矩阵为

$$\frac{\partial^2 l(\beta)}{\partial \beta_j^2} = -\sum_{i=1}^{n} x_{ij}^2 \pi_i (1 - \pi_i) \tag{3.9}$$

$$\frac{\partial^2 l(\beta)}{\partial \beta_j \partial \beta_i} = -\sum_{i=1}^{n} x_{ij} x_{ji} \pi_i (1 - \pi_i) \tag{3.10}$$

若写成矩阵形式，以H表示Hessian矩阵，X与C分别表示

$$X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix} \tag{3.11}$$

$$C = \begin{bmatrix} \hat{\pi}_1(1-\hat{\pi}_1) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \hat{\pi}_n(1-\hat{\pi}_n) \end{bmatrix} \tag{3.12}$$

则$H = X^T CX$，再令

$$V = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix} * \begin{bmatrix} y_1 - \pi_1 \\ \vdots \\ y_p - \pi_p \end{bmatrix} \tag{3.13}$$

即为似然方程的矩阵形式。

在此基础上得到了牛顿迭代法的形式为

$$W_{new} = W_{old} - H^{-1}V \tag{3.14}$$

(3.14)式中H为对称正定的，求解$H^{-1}V$即为求解线性方程$HX = V$中的矩阵X。

最大似估计的渐近方差(asymptotic variance)与协方差(covariance)可由信息矩阵(information matrix)的逆矩阵估计出来。

第三节　显著性检验

一、Wald 检验

针对回归系数进行显著性检验时，通常会使用 Wald 检验，其公式为

$$W = [\hat{\beta}_j / \widehat{SE}(\hat{\beta}_j)]^2 \tag{3.15}$$

其中，$\widehat{SE}(\hat{\beta}_j)$为$\hat{\beta}_j$的标准误差估计值。这个单变量 Wald 统计量服从自由度为1的χ^2分布。如果需要检验假设$H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0$，则计算统计量为

$$W = [\hat{\beta}' / \widehat{SE}(\hat{\beta}')]^2 \tag{3.16}$$

其中，$\hat{\beta}'$为去掉$\hat{\beta}_0$所在的行和列的估计值，相应地$\widehat{SE}(\hat{\beta}')$为去掉$\hat{\beta}_0$所在的行

和列的标准误差的估计值，这里 Wald 统计量服从自由度为2的χ^2分布。

然而当回归系数的绝对值很大时，这一系数标准误的估计值会膨胀，于是导致 Wald 统计值变得很小，以至于第二类错误的概率增加。因此当发现回归系数的绝对值很大时，就不再用 Wald 统计值来检验零假设，此时应该使用似然比(likelihood ratio)检验来替代。

二、似然比检验

在一个模型中，含有变量x_i与不含有该变量的对数似然值乘以-2的结果之差服从卡方分布。这一检验统计量被称为似然比(likelihood ratio)，用式子表示为

$$G = -2\ln\frac{\text{不含}x_i\text{的似然}}{\text{含有}x_i\text{的似然}} \tag{3.17}$$

计算似然值采用公式(3.6)，倘若需要检验假设$H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0$，则计算统计量为

$$G = 2\left[\sum_{i=1}^{n} y_i \ln(\hat{\pi}_i) + (1-y_i)\ln(1-\hat{\pi}_i)\right] - [n_1 \ln(n_1) + n_0 \ln(n_0) - n\ln(n)] \tag{3.18}$$

在(3.18)式中，n_0表示$y_i = 0$的观测值的个数，n_1表示$y_i = 1$的观测值的个数，则n就表示所有观测值的个数。实际上，(3.18)式等号右端的右半部分$[n_1 \ln(n_1) + n_0 \ln(n_0) - n\ln(n)]$表示只含有$\beta_0$的似然值，并且统计量$G$服从自由度为$p$的卡方分布。

三、得分检验

在零假设$H_0: \beta_k = 0$下，假设参数的估计值为$\beta_{(0)}$，则计算 Score 统计量的公式为

$$U(\beta_{(0)})^T I^{-1}(\beta_{(0)}) U(\beta_{(0)}) \tag{3.19}$$

在(3.19)式中，$U(\beta_{(0)})$表示在H_0下的对数似然函数(3.7)与(3.8)的一阶偏导数值，而$I(\beta_{(0)})$则表示在H_0下的对数似然函数(3.7)与(3.8)的一阶偏导数值，并且

Score 统计量服从自由度为1的卡方分布。

第四节　模型扩展

在之前介绍的多项或者二项 Logistic 回归分析模型的基础上，本节着重于介绍模型的扩展。该想法主要来源于一个考虑：互联网财经新闻对于股票市场价格波动的影响可能存在滞后性，同时也不排除股价波动会对本身有一定的影响。这样就使得模型发生了改变，即模型中y_t的取值将受到除原模型中的自变量x_t之外，还可能有当前时刻t之前时刻的响应变量y_{t-1}^*，其中$y_{t-1}^* = (y_{t-1}, \dots, y_1)$。

因此模型中的设计矩阵将变为

$$Z_t = (x_t, y_{t-1}^*)$$

针对响应变量为二元时间序列的情形，COX(1970)提出了 autoregressive logistic 模型，其中协变量与有限数量过去时间的输出值作为线性预测的一个组成部分。令

$$H_t = \{y_{t-1}, y_{t-2}, \dots, y_1, x_t, x_{t-1}, \dots, x_1\}$$

代表过去观测值、现在与过去协变量的历史集合。则广义自回归模型具有以下结构特点：

(1)条件密度函数$f(y_t|H_t)$, $t = 1,2,\dots$是指数分布族

(2)条件期望$\mu_t = E(y_t|H_t)$的形式如下

$$\mu_t = h(z_t'\beta) \tag{3.20}$$

其中是h广义自回归模型中对应的响应函数，p维设计向量z_t是关于H_t的函数，即$z_t = z_t(H_t)$。

假定(1)与(2)暗含着条件方差$\sigma_t^2 = var(y_t|H_t)$的形式为：

$$\sigma_t^2 = \nu(\mu_t)\phi \tag{3.21}$$

其中$\nu(\cdot)$对应特定的指数分布族中的方差函数，ϕ是尺度参数。一般来说，设计向量z_t是由响应过程与（或）任何与响应变量同时相关的协变量的任意数量滞后值构成。但是通常只有有限数量过去的观测值$y_{t-1}, \dots, y_{t-l}$会被考虑在设计向量中，这种模型称之为广义自回归模型或者l阶马尔科夫模型。

针对本文讨论的问题，响应变量为二项时间序列$\{y_t\}, y_t \in \{0,1\}$，在给定H_t下

y_t的条件分布由：

$$\pi_t = P(y_t = 1|H_t) \tag{3.22}$$

所决定，在包含协变量的情况下有

$$\pi_t = h(\beta_0 + \beta_1 y_{t-1} + \cdots + \beta_l y_{t-l} + x_t'\gamma) = h(z_t'\beta), t > l$$
$$z_t' = (1, y_{t-1}, \ldots, y_{t-l}, x_t'), \beta' = (\beta_0, \ldots, \beta_l, \gamma'), \tag{3.23}$$

以$l = 1$为例，$\pi_t = h(\beta_0 + \beta_1 y_{t-1} + \beta_2 x_t)$等价于以下参数化的转移概率：

$$\pi_{i0} = P(y_t = 1|y_{t-1} = i, x_t) = h(\alpha_{0i} + \alpha_{1i} x_t), i = 0,1$$
$$\pi_{i1} = P(y_t = 1|y_{t-1} = i, x_t) = 1 - \pi_{i0} \tag{3.24}$$

令$y_{t-1} = 0$或1可以很容易得到

$$\alpha_{00} = \beta_0，\ \alpha_{01} = \beta_0 + \beta_1，\ \alpha_{10} = \alpha_{11} = \beta_2$$

第五节　估计方法

本节将针对上一节给出的模型扩展形式提供估计方法，广义自回归模型的估计与检验都以真正的似然模型为基础。在确定性协变量的情形中，$y_1, \ldots, y_T$的联合密度可以因式分解为条件密度的乘积，即

$$f(y_1, \ldots, y_T|\beta) = \prod_{t=1}^{T} f(y_t|y_{t-1}, \ldots, y_1; \beta), \tag{3.25}$$

其中条件密度取决于假定一些特定的广义自回归模型。如果协变量是随机的，则联合密度分解为

$$f(y_1, \ldots, y_T, x_1, \ldots, x_T|\beta) = \prod_{t=1}^{T} f(y_t|H_t; \beta) \prod_{t=1}^{T} f(x_t|C_t), \tag{3.26}$$

其中$C_t = (x_1, \ldots, x_{t-1})$。假定(3.26)式中的第二个乘积项不依赖于$\beta$，则估计可以由以第一个乘积定义的部分似然为基础得到。在任何情形中，$y_1, \ldots, y_T$的对数似然为

$$l(\beta)=\sum_{t=1}^{T} l_t(\beta),\ l_t(\beta)=\log f(y_t|H_t;\beta),$$

(3.27)

其中的条件密度由广义自回归模型的定义给定。对l阶马尔科夫模型，严格来说这是一个给定初值$y_0,\dots,y_{-l+1}$下的条件对数似然。得分函数即$l(\beta)$的一阶导数为：

$$s(\beta)=\sum_{t=1}^{T} Z_t' D_t(\beta)\Sigma_t^{-1}(\beta)(y_t-\mu_t(\beta))$$

(3.28)

其中$\mu_t(\beta)=h(Z_t(\beta))$是条件期望，$\Sigma_t(\beta)=\mathrm{cov}(y_t|H_t)$是条件方差矩阵，$D_t(\beta)=\partial h/\partial\eta$是在$\eta_t=Z_t\beta$时取值。对单元响应变量，$Z_t$降低为$z_t'$，$\Sigma_t(\beta)$则为条件方差$\sigma_t^2$。从形式上看，这与把过去的响应变量看成额外增加的协变量是相同的。但是写出无条件期望信息矩阵$F(\beta)=EF_{obs}(\beta)$的显示解通常是不可能的。相反的，对相依观测值来说条件信息矩阵为：

$$G(\beta)=\sum_{t=1}^{T}\mathrm{cov}(s_t(\beta)|H_t)$$

(3.29)

其中$s_t(\beta)=\partial\log f(y_t|H_t;\beta)/\partial\beta=Z_t'D_t(\beta)\Sigma_t^{-1}(\beta)(y_t-\mu_t(\beta))$是个别得分函数，在计算与渐近考虑中发挥一个重要的作用。因此条件信息矩阵化为

$$G(\beta)=\sum_{t=1}^{T} Z_t' D_t(\beta)\Sigma_t^{-1} D_t'(\beta)Z_t$$

(3.30)

且在形式上与独立观测的期望信息矩阵相同。整合观测值$\{y_t\}$后原则上可以得到无条件期望信息$F(\beta)=\mathrm{cov}\,s(\beta)$。

为了计算$l(\beta)$对应（局部）最大值$s(\beta)$的极大似然估计$\hat{\beta}$，可以将过去响应变量值对待为额外增加的协变量而采用与独立观测相同的迭代算法。即可将(3.30)式替代得分算法或者等价的迭代加权最小二乘估计过程中的无条件期望。在合适的“正则假定”条件下，最大似然估计$\hat{\beta}$是一致和渐近正态的：

$$\hat{\beta}\sim N(\beta,G^{-1}(\hat{\beta}))$$

(3.31)

其中条件信息矩阵的逆为该渐近分布的协方差矩阵。由于相依观测相对独立观测包含更少的信息，因此这类的正则假定通常会有更多限定。通常平稳性与遍历性的假定条件是为了让随机变量的平稳序列可以适用极限理论。某些非平稳的形式可以采用此方法，但是需要增加额外的数学研究。为了说明哪些类型的非平稳可以适用，本文将以(3.20)式且设计向量为$z_t' = (1, y_{t-1}, \dots, y_{t-l}, x_t')$的自回归逻辑模型为例进行讨论。因此以下两个条件对于确保最大似然估计的一致性与渐近正态性是足够的：

(1)协变量序列$\{x_t\}$是有界的；

(2)经验协方差矩阵

$$S_t = \sum_{s=1}^{t} (x_s - \bar{x})(x_s - \bar{x})'$$

是发散的，如$\lambda_{min} S_t \to \infty$或者等价的$S_t^{-1} \to 0$。

如果用条件信息矩阵替代非条件信息矩阵以此完全类似于本章第三节的方法利用似然比、Wald 和得分检验验证线性假定是可能的。同时本质上在最大似然估计的一致与渐近正态性所需的相同条件满足下检验统计量一般的渐近性质依然有效。

第四章　互联网财经新闻对股票的影响

在前几章中分别介绍了论文的研究背景和相关理论知识等工作，本章节将介绍本文的核心部分，包括数据准备、模型建立以及分析模块。

第一节　研究样本及思路

本文研究所需要的数据量比较大，因此利用本节来介绍研究所需要的数据的准备以及基本分析工作。本研究会使用到的数据包括互联网财经新闻和股票市场的交易日数据，本章节将会介绍数据的获取、预处理以及基本分析。对于学统计的人来说，数据预处理是数据准备中十分重要的任务，这也会直接影响之后情感分析和建立模型等的结果。

一、数据获取

由于本论文旨在研究目标企业在一段时间内出现在目标新闻源上的个股新闻所能传达给读者的情感倾向，以及这些情感倾向会对股票市场价格的波动产生怎样的影响。故本论文采用专门的网页爬虫技术，选取了新浪财经作为新闻源，自动抓取所研究的企业在目标时间段内的新闻文本标题。

本文选择的新闻来源是“新浪财经”(www.finance.sina.com.cn/)。首先是因为新浪财经拥有超大的用户量以及访问量；再者新浪财经以其财经资讯的广度与深度受到金融机构和投资者的信赖，在中国金融行业有很大的影响。本文选择的目标企业是不同类型的五家银行为研究主体，分别为：中国银行(601988)、农业银行(601288)、交通银行(601328)、招商银行(600036)以及浦发银行(600000)。选择这五家上市公司作为研究主体的原因：（1）本文选择同一行业的上市公司是希望剔除行业因素的影响；（2）希望通过不同类型的上市银行来探讨所尝试研究的问题的实用性。而考虑到最近 2015 年 6 月底发生过的最严重的一次股灾，考虑到这次股灾造成一定中小股民的资产蒸发以及股民心态的转变，因此本文研究的时间窗口为股灾发生前的 2014 年 5 月 1 号到 2015 年 6 月 9 号，其中时间窗口为 2015 年 3 月 1 号到 2015 年 6 月 9 号之间的数据将用于模型检验。研

究窗口选定之后，接下来就是通过爬虫软件从目标新闻源上抓取相应的财经新闻。通过上一章节的介绍，本文针对所要研究的新闻源，利用 R 批量抓取关于中国银行等五家银行在研究时间窗口内出现在网页上的所有个股新闻标题。相应的程序见图 4.1

```
library(RCurl)
library(XML)
library(plyr)


####给定所有网页，提取此新闻网的一段时间内的新闻
getnews <- function(URL){
  stockinfs = laply(URL,function(URL1){
    temp <- getURL(URL1,.encoding="gb2312")
    temp1<-iconv(temp,"gb2312","UTF-8") #转码
    Encoding(temp1) #UTF-8
    Sys.sleep(runif(1,1,2))
    doc <- htmlParse(temp1,encoding="UTF-8")
    rootNode <- xmlRoot(doc)
    stockinf <- xpathSApply(rootNode,"//div[@class='datelist']/ul/a",xmlValue)
    return(stockinf)
  })
  return (stockinfs)
}
```

图 4.1 新浪财经新闻抓取函数

通过上段程序可以得到新浪财经相应页面上的所有关于中国银行等五家银行的个股新闻，总共有 2327 条新闻。可以通过表 4.1 查看其中一小部分从新浪财经上抓取下来的新闻标题以及对应的日期。

表 4.1 新浪财经抓取的部分新闻样本

新浪财经新闻标题	日期
银行股王者归来 宁波南京银行涨停	2015-03-09
混业经营引爆银行股 金融控股模式成大方向	2015-03-09
银行业:金改趋势不变 业绩短期底部将出现	2015-03-09
逾 55 亿昨疯抢银行股 三大维度破解大涨玄机	2015-03-09
中国银行关于“中行转债”赎回结果及兑付摘牌的公告	2015-03-10

其他四家银行只需改变对应的股票代号以及页面，即可得到不同银行研究窗口内在新浪财经上的所有个股新闻标题，最终关于农业银行(601288)的新闻标题有 1907 条，交通银行(601328)有 1900 条，招商银行(600036)有 1940 条以及浦发银行(600000)有 1760 条。

本文所用的股票市场交易数据来源于 Wind 资讯金融终端，这是一个在线实时金融信息终端，为用户提供行情报价、金融数据和财经信息等功能的综合性

金融服务平台。利用此平台可以获得任何上市公司股票在交易期内的价格波动变化，因此针对本文研究的股票（以中国银行 601988 为例）以及研究窗口 2014 年 4 月 1 号到 2015 年 6 月 9 号，可以得到该股票价格的波动，部分价格变化如表 4.2 所示

表 4.2 中国银行部分股市交易数据

时间	开盘	最高	最低	收盘	涨幅
2015-03-06,五	3.78	3.88	3.77	3.83	0.79%
2015-03-09,一	3.85	4.06	3.81	4.03	5.22%
2015-03-10,二	3.98	4	3.92	3.93	-2.48%
2015-03-11,三	3.95	4.08	3.95	4.01	2.04%
2015-03-12,四	4.09	4.26	4.07	4.18	4.24%

二、研究思路及目的

（一）整体思路

本论文将针对中国银行(601988)在新浪财经等新闻源上在研究窗口 2014 年 5 月 1 号到 2015 年 6 月 9 号之间的个股新闻标题进行情感分析，其中时间窗口为 2015 年 3 月 1 号到 2015 年 6 月 9 号之间的数据将用于模型检验。再利用 Logistic 回归分析模型研究这些个股新闻对股票价格波动趋势的影响。该分析主要包括两方面。

（1）利用文本挖掘技术以及情感分析对所得到的关于中国银行(601988)的个股新闻标题进行分词以及情感分析，得到每条新闻标题的情感得分并进行一定的处理；

（2）应用上一步骤得到的每只股票每天的情感得分以及对应股票市场价格变化的波动情况，分别进行 Logistic 回归分析并进行预测。

（二）研究目的

本章的研究目的是较为系统的将文本挖掘与 Logistic 回归分析的研究方法引入到对互联网财经新闻标题对于股票价格变动趋势的影响并用于预测，在此基础上为投资者以及企业管理者提出一些针对性的建议。

第二节　互联网财经新闻的情感分析

本节着重于针对关于中国银行(601988)等五家银行在新浪财经等新闻源上获取的新闻标题进行情感倾向分析并给予相应的得分，主要会利用到的编程语言是 R。

一、对新闻进行中文分词

在进行情感分析过程中，首先要做的是对每一句文本进行中文分词。分词过程主要使用的软件包是 Rwordseg，在进行分词的时候除了利用该软件包内含的中文分词算法，可以利用搜狗拼音中提供的多个词库更新丰富已有的分词词库。因为分词的准确程度会直接影响之后的情感倾向的判断以及情感得分的评定。由于本论文讨论研究的是关于财经新闻对于股价变动的影响，因此后期加入的词库也偏向于财经等相关领域，包括有财经金融、股票、数字时间和证券等词汇。

与英文每个单词之间通过空格自然分开不同，中文最小的单位只是字，某个字与前后哪些字合在一起组成一个词并不好定义。例如，英文句子 I like sunshine，中文翻译则为：“我喜欢阳光”。计算机可以自动的通过空格将 sunshine 看成一个单词，但是却并不一定可以将于“阳”与“光”两个字连在一起看成一个词。而为了将汉字按照一定的规则分割成有意义的词就是中文分词。

除了中文分词本身的困难，中文的词还会随着时间与网络发展等因素发生一定的改变与更新，这其中也包括例如公司名称等词汇的增加。以“淘宝”为例，在载入搜狗拼音词库之前，利用软件分词得到的结果是“淘”与“宝”，这样的分词结果明显影响源文本句子所要表达的信息。

除了考虑新的词汇等影响因素外，针对中文的分词效果还会受到停用词(stop words)的影响。所谓停用词，即是指为节省存储空间和提高搜索速度与效率，搜索引擎在索引页面时会自动忽略的某些字或词。通常停用词大致可以分为以下两类：

(1) 使用十分广泛甚至过于频繁的单词或者是字。比如中文中的“就”、“是”，英文中的“is”、“how”等之类几乎会在每个文档中出现的词，查询这类词搜索引擎并不能保证得到真正相关并有意义的结果，难以提高搜索结果的准确性的同时还会大大降低搜索的效率。

(2) 在文本中的实际意义并不大但出现频率很高的词。这一类词主要包括有副词、介词、语气助词和连词等，通常情况下这些词本身并无明确的意义，只有将这些词放入一个完整的句子中才会起到作用。例如中文的“的”、“和”、“在”之类的词。

文档中如果大量使用停用词将会对其有效信息产生干扰。因此，在进行情感分析时，本文会在分词之后对那些停用词进行降噪处理，即将会对情感得分产生影响的停用词消除掉。具体的程序见图 4.2：

```
####### clean up sentences with R's regex-driven global substitute, gsub():
sentence = gsub('[[:punct:]]', '', sentence)
sentence = gsub('[[:cntrl:]]', '', sentence)
sentence = gsub('\\d+', '', sentence)

####### split into words
word.list = lapply(X = sentence,FUN = segmentCN)
####### sometimes a list() is one level of hierarchy too much
words = unlist(word.list)

#######去掉停用词
data_stw=read.table(file="stopwords.txt",colClasses="character")
stopwords_CN=c(NULL)
for(i in 1:dim(data_stw)[1]){
  stopwords_CN=c(stopwords_CN,data_stw[i,1])
}
for(j in 1:length(stopwords_CN)){
  words <- subset(words,words!=stopwords_CN[j])
}
```

图 4.2 中文分词程序

在图 4.2 中的中文分词程序中，首先针对输入的 sentence 的一些符号例如空格，逗号等剔除，因为这些符号基本不会带来信息。之后在此基础上利用函数 segmentCN 进行分割，针对此函数进行一定的介绍

```
segmentCN(strwords,
  analyzer = get("Analyzer", envir = .RwordsegEnv),
  nature = FALSE, nosymbol = TRUE,
  returnType = c("vector", "tm"), isfast = FALSE,
  outfile = "", blocklines = 1000)
```

图 4.3 segmentCN 函数介绍

处理对象： strwords 可以是一段中文，或是某个文本文件的路径，并且可用 outfile 参数指输出，默认是原路径下；词性输出：参数 nature 可以设置是否输出词性，默认不输出，如果选择输出，那么返回的向量名为词性的标识；输出内容：参数 nosymbol 表示是否只输出汉字、英文和数字，默认为 TRUE，否则将会输入标点符号等。部分的分词结果显示如表 4.3 所示

表 4.3 中文分词部分结果

新闻标题	分词结果
快讯中字头股集体暴动 中国远洋等股涨停	“快讯”“中字头”“股”“集体”“暴动”“中国远洋”“股”“涨停”
传五银行将在上海自贸区对境外发行同业存单	“传”“五”“银行”“上海”“自贸区”“境外发行”“同业”“存单”
快讯银行股集体暴涨 中国银行等股涨停	“快讯”“银行股”“集体”“暴涨”“中国银行”“股”“涨停”
混改预期引爆银行股 浦发停牌交行再度涨停	“混改”“预期”“引爆”“银行股”“浦发”“停牌”“交行”“再度”“涨停”

从上表可以看出利用此函数得到的分词效果基本上可以认为是较为合乎现实理解的，首先将一些公司词汇如“中国远洋”、“中国银行”等分为一个词；其次也能够将股票、财经等上专有的词汇如“涨停”、“同业”等词汇划分为一个词汇。

二、计算情感得分

在上一节得到的中文分词基础上，通过第二章中介绍的中文情感分析方法进行情感分析。本文使用的情感分析方法是基于情感词典的方法，该方法的第一步是确定一个词是积极还是消极，这一步的结果好坏直接取决于词典的完整程度。英文在这方面的研究比较先进完善，已经有伟大的词典资源 SentiWordNet。在这部词典中无论是积极、消极或主观、客观亦或者是词语的情感强度值都可以得到。但是，在中文领域这方面的词典的实用效果并不甚好，其主要原因在于词典资源质量不高以及中文不同词的发展比较快。因此在这方面除了利用得到的一些中文词典外，考虑到研究的主体是财经新闻，本文手动加入了一些在股票财经等方面会使用到的能够表达一定情感倾向的词汇。除此之外，还需要考虑程度副词对整个文档情感得分的影响。例如“很开心”所表达的情感是积极的，并且积极程度明显比“开心”更深。因此还需要一个较为完整的程度副词词典，同时针对不同程度的副词给予相应不同的值。

在得到了比较完整的情感词典以及程度副词后，情感分析需要进行第二步，即识别一个句子是积极还是消极的。主要的方法是将文档分词后的每个词通过自行编写的搜索函数在给定的情感词典以及程度副词中查询，进行词表匹配，

自定义一个函数 mysearch。该函数的目的为若某个词在词表中存在，返回其所在位置；否则返回 NA。主要需要针对每一个文档得到的分词结果进行一定的搜索。具体函数见图 4.4：

```
# mysearch() returns the position of the matched term or NA
mysearch <- function(x,y){
  n = length(y)
  m = rep(0,n)
  for(i in 1:n){
    if (length(which(x==y[i])) ==0)
      m[i] = NA
    else m[i] = which(x==y[i])
  }
  return(m)
}
```

图 4.4 中文分词词表匹配函数

利用函数 mysearch 得到积极、消极和程度副词的词表匹配结果

```
# compare our words to the dictionaries of positive negative & degree terms
pos.matches = mysearch(pos.words,words)
neg.matches = mysearch(neg.words,words)
de.matches = mysearch(de.words,words)
```

图 4.5 不同词的匹配函数

在得到相应的匹配词后，需要进一步考虑程度副词的影响，本文只考虑程度副词之后直接跟着情绪词的情况。如果情感词之前直接有程度副词修饰，则先将积极和消极匹配词变为 0/1，然后再根据程度副词给定的相应值进行查找与替换。最后将得分比上句子的长度得到最终情感得分

```
pos.matches = as.numeric(!is.na(pos.matches))
neg.matches = as.numeric(!is.na(neg.matches))

n= length(de.matches)
for (i in 1:(n-1)){
  if (!is.na(de.matches[i])) {
    de = de.score[de.matches[i]]
    if (pos.matches[i+1]==1)
      pos.matches[i+1] = de
    else if (neg.matches[i+1]==1)
      neg.matches[i+1] = de
  }
}

score = sum(pos.matches) - sum(neg.matches)
score = score/length(words)
```

图 4.6 中文情感分析最终得分

利用图 4.5 与图 4.6 中显示的函数可以得到每个文档对应的情感得分，一部分情感得分样本如表 4.4 所示：

表 4.4 中文情感得分部分样本

新闻文档	情感得分
张江高科与中行合作 895 创业营或成 P2B 众筹模式首单	0.272727273
交行混改方案拟引入民营资本 触及两项“敏感点”	-0.181818182
银行行业:中行重组东盟地区业务 大行争相拓展海外版图	0.25
详解深证成指大扩容的投资机会	0
金融业:银行非经济低迷之祸首	-0.2

以文档“银行行业:中行重组东盟地区业务 大行争相拓展海外版图”为例，利用分词函数得到的分词结果为“银行”、“行业”、“中行”、“重组”、“东盟”、“地区”、“业务”、“大行”、“争相”、“拓展”、“海外”和“版图”，对应的长度为 12。然后利用不同词的匹配函数图 4.4 与图 4.5 得到该句子中带有情感倾向的词为“重组”、“争相”和“拓展”，这些词的出现往往代表企业有一定的推动公司发展的行为或者举措，因此这些都属于积极情感倾向得词汇；同时在这句文档中并没有表示程度的副词，因此根据第二章中提到的关于情感分析的理论知最后的情感得分为 score=3/12=0.25。

此时得到的是每一条新闻标题对应的情感得分，而实际上关于研究主体以中国银行(601988)为例的银行企业在一天之内出现在新浪财经上经常会可能有多条新闻。因此本文需要对最初得到的每条新闻的情感得分进行一定的处理以此得到建模需要的自变量。

首先，针对一天有多条新闻的情况本文采用的处理方式是：将同一天在一个新闻网上出现的新闻情感得分相加得到这一天该新闻网的情感得分。简单来说该处理方式的想法在于如果是同一倾向的新闻显然会增强读者对该股票在此倾向上的认知；而如果两个相反情感倾向的新闻同时存在，则读者会被相反的新闻减弱本来已有的情感倾向；

其次，通过表 4.2 可以看出由于股票市场并不是每一天都有交易，最明显的是周末两天股票交易所是休息的因此没有交易数据。此外，如果遇到企业开股东大会或者有重大事件等都会停牌，这时候即便是正常交易日也可能没有交易数据。因此，如果遇到没有交易数据的情况，则这几天内发生的新闻将对紧接

下来有交易数据日的新闻情感得分造成一定的影响。

针对这两种情况本文的处理方式是：以“天”为单位，加入股市交易作为考察基准。正常交易日（前一天也为交易日）的新闻作为交易日当天的数据，而非交易日发生的新闻则视为会对最近一个交易日产生影响，例如周末发生的新闻会影响下周一股市的价格波动。当新闻发布日(t)对应的股市无交易时，该新闻日对应情感得分输出为第二天(t+1)的股价波动；若第二天股市仍无交易，则对应为第三天(t+2)的股价波动。以此类推，但是考虑到新闻的时效性，此处理方法不超过(t+3)，若超过则放弃这则新闻。此外，新闻时效性也决定了交易日当天的新闻影响更大，因此处理过程还加入了加权平均等方法。

通过上述介绍的关于情感得分预处理可以看到，处理过程中大多时候会遇到一天内发生多条新闻，而最终的情感得分并没有直接显示出这一信息，这一处理有一定的信息遗漏。因此本文将每天发生的新闻条数作为另一个变量，称为热度。

最终得到研究窗口对应的股票交易日内不同研究主体对应的所有情感得分：

表 4.5 部分时间内各银行情感得分

日期	中国银行	农业银行	交通银行	招商银行	浦发银行
2014/9/1	1.261869	0.892551	0.504293	0.640505	0.256818
2014/9/2	0.775433	0.334066	0.665909	0.619048	0.142857
2014/9/3	0.509524	0.993290	0.609524	0.384524	0.432601
2014/9/4	0.359524	0.276190	0.045421	0.776190	0.476190
2014/9/5	0.100000	0.441964	0.285714	0.687637	0.285714
2014/9/9	0.769444	0.472222	0.236111	0.071429	0.009091
2014/9/10	-0.078438	-0.160256	0.147752	-0.146245	-0.328438
2014/9/11	0.055556	0.076923	0.000000	0.017399	0.000000
2014/9/12	0.020202	0.111111	0.461111	0.111111	0.290598
2014/9/15	0.756746	0.280952	0.488889	0.313131	0.222222
2014/9/16	0.780664	0.249928	1.128571	0.142857	1.020635

第三节　新闻与股市价格波动的实证研究

本节着重于针对关于中国银行(601988)等五家银行在新浪财经等新闻源上的

新闻标题情感得分与相应的股票市场价格波动情况进行建模分析。首先针对每个上市公司的情感得分、新闻热度以及其他可能会对股市价格造成一定影响的解释变量进行基本的 Logistic 回归分析；在此基础上将进一步考虑前一个交易日响应变量与自变量共同在模型中现交易日响应变量产生影响，再次进行 Logistic 回归分析并进行预测，这过程中主要会利用到的编程语言是 R。

通常情况下，一个上市公司的股票价格会受到很多因素的影响，总的来说大概可以分为以下几个大的方面：

（1）公司内部因素。这个方面包括有公司的经营状况、盈利能力以及财务状况等，这是影响股票价格最重要的基本意思。因为这些因素直接反映公司的各种状况，亦可间接影响投资者的投资意向，进而影响股价的变化。当公司的经营状况良好，盈利能力强时，公司的股票价格基础扎实，股民自然认为该公司的股票很稳定，上涨的机会就多；

（2）宏观经济因素。包括经济周期等，其中最近的 2015 年 6 月底发生股灾就是很好的关于经济周期对于股价影响的例子。股价总是伴随着经济周期的变化而升降，例如在经济复苏阶段，投资逐步回升，资本周转开始加速，利润逐渐增加，股价呈上升趋势。而遇到经济危机时，由于支付能力的需求减少，造成整个社会的生产过剩，企业经营规模缩小，产量下降，失业人数迅速增加，企业的盈利能力急剧下降，股价自然随之下跌；

（3）政策因素。在这方面包括的因素比较多，可能有政治因素、经济政策、利率、税收制度和信用政策等。考虑经济政策，例如当国家对某些行业或某类企业增加投入，就意味着这些行业、企业的生产将发展，亦会同样引起投资者的重视，股价将随之上涨。而利率一般与股价成反比，利率的上升将导致公司借款成本增加以及资金从股市流入银行等，这将导致股票价值下降。

由于本文研究的是交易日内互联网财经新闻对于股票价格波动的影响，因此考虑的自变量需要时以“天”研究单位。故针对以上对企业股票价格会产生影响的四类因素，本文将选择几个有一定代表性且以“天”为研究单位的变量，以及上一节得到的交易日期内每个新闻源对应的情感得分与新闻热度值的对数值共同作为各自 Logistic 回归分析的自变量。本文将选择的变量包括：隔夜拆借利率(SHIBOR %)、即期汇率、上证指数以及行业指数。其中，隔夜拆借利率代表的是会对其他利率产生影响的政策因素且每个交易日都会有一个报价；选择即期汇率是因为认为该变量会影响股票市场资本流出和股市回调，代表中国宏

观经济因素；选择上证指数是因为本论文考察的五家上市公司都是在沪市上市的，会受到大盘指数的影响；最后选择行业指数是因为这五家上市公司同属于银行行业，故可以将行业指数加入模型进行考虑。至于没有从公司内部因素中选择变量是因为响应变量是以“天”为研究单位，而公司内部因素中很难找到满足这一条件的变量，因此本文暂时不考虑将这一因素放入模型中。

一、Logistic 回归分析

由于本文的研究目标因变量是股票价格波动的情况，该变量取决于交易日收盘价与开盘价之间的涨跌幅情况。对于股民而言比较关心的一直都是股票发生涨幅的情况，因此针对涨跌幅情况本文的处理方式为：简单的将涨跌幅情况分为两个值，当涨幅大于 0 时，则因变量取值为 1；否则，遇到其他情况都认为不存在涨幅，则因变量取值为 0，即发生涨幅(Y=1)与不发生涨幅(Y=0)。在本文讨论的六个解释变量同时存在的情况下得到的二项 Logistic 回归分析模型为：

$$\ln\left(\frac{P(\mathrm{Y}=1)}{P(Y=0)}\right)=\beta_0+\beta_1 score+\beta_2 \text{汇率}+\beta_3 \mathrm{SHI}+\beta_4 \ln \mathrm{ba}+\beta_4 \ln \mathrm{sz}+\beta_5 \ln \text{热度} \tag{4.1}$$

Score 代表新闻情感得分，对变量“银行指数 ba”、“上证指数 sz”和“热度”取对数是为了减少或者消除异方差的存在。

（1）模型估计

首次建模时将六个解释变量全部放入模型中，以此得到五家银行新闻对其股票价格波动趋势影响的统计分析结果如下表 4.6 所示。其中本文采用了似然比检验来保证模型的显著性，似然比检验结果见于表 4.6 最后一列。

表 4.6 初步 Logistic 回归分析结果

	截距	Score	SHI	lnbank	lnsz	汇率	ln 热度	LRT
中	80.822 (0.021)	0.563 (0.041)	0.487 (0.436)	16.058 (0.002)	-15.566 (0.003)	-10.291 (0.047)	0.420 (0.216)	22.830 (0.001)
农	88.426 (0.015)	0.506 (0.124)	0.535 (0.428)	13.448 (0.012)	-11.059 (0.035)	-14.453 (0.010)	0.049 (0.827)	16.580 (0.011)
交	93.967 (0.011)	0.917 (0.014)	0.525 (0.434)	15.843 (0.003)	-15.808 (0.003)	-11.732 (0.037)	-0.145 (0.609)	19.930 (0.003)
招	15.978 (0.625)	1.654 (0.000)	0.905 (0.131)	4.522 (0.378)	-4.960 (0.318)	-1.382 (0.773)	-0.302 (0.325)	25.480 (0.000)
浦	62.007 (0.058)	1.297 (0.003)	0.711 (0.216)	6.920 (0.174)	-6.900 (0.163)	-8.775 (0.063)	-0.064 (0.794)	20.100 (0.003)

注：表中括号上面的数是 Logistic 回归系数的参数估计值，括号里面的数为回归系数 t 检验对应的 p 值。

从表 4.6 中可以看出，对于每个银行而言在模型中 Score 这个变量在 10%的水平下全部都显著；且每个银行对应的模型在置信水平 0.05 下整体都是显著的。然而可以看出每个银行得到的模型中都存在一些变量并不显著的情况，因此下一步将对模型中的解释变量进行一定的删减，尽量在保留足够变量下使得模型中所有的解释变量在 0.05 置信水平下达到显著，则相应的统计分析结果如表 4.7 所示：

表 4.7 Logistic 回归分析最终结果

	截距	Score	lnbank	lnsz	汇率	SHI	LRT
中	81.937 (0.020)	0.754 (0.005)	16.629 (0.002)	-15.879 (0.002)	-10.312 (0.043)	*	20.680 (0.000)
农	86.034 (0.014)	0.543 (0.047)	14.614 (0.004)	-12.313 (0.014)	-13.418 (0.010)	*	15.940 (0.003)
交	86.120 (0.017)	0.807 (0.008)	16.459 (0.002)	-16.602 (0.001)	-9.879 (0.061)	*	18.940 (0.001)
招	-3.496 (0.023)	1.401 (0.000)	*	*	*	1.068 (0.046)	22.940 (1e-05)
浦	60.701 (0.066)	1.214 (0.001)	8.180 (0.100)	-8.080 (0.097)	-8.050 (0.086)	*	18.340 (0.001)

注：*号代表该解释变量在此银行对应的模型中被剔除了，括号上面的数是 Logistic 回归系数的参数估计值，括号里面的数为回归系数 t 检验对应的 p 值。

且该表得到的最终模型中大多数解释变量在 0.05 置信水平下达到显著。

（2）模型解释

根据表 4.7，以中国银行为例最终得到的二项 Logistic 回归分析模型为：

$$\ln\left(\frac{P(\mathrm{Y}=1)}{P(Y=0)}\right)=81.937+0.754sc+16.629\mathrm{lnba}-15.879\mathrm{lnsz}-10.312\text{ 汇率} \tag{4.2}$$

由模型式可知，若情感得分 score 增加 1 分，股票价格产生涨幅的优势比为：$\exp(\hat{\beta}_1)=\exp(0.754)=2.125$即情感得分 Score 增加 1 分，股票价格产生涨幅的优势增加 100.125%。

其中该模型对应的参数估计结果显示如图 4.7 所示：

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  81.9367    35.3307   2.319  0.02039 *
score         0.7543     0.2704   2.790  0.00527 **
lnbank       16.6286     5.2438   3.171  0.00152 **
lnsz        -15.8793     5.1431  -3.087  0.00202 **
即期汇率    -10.3121     5.0871  -2.027  0.04265 *
---
```

图 4.7 中国银行估计结果

因此对参数β_1 (Score) 95%的置信区间是：

$$\hat{\beta}_1\pm z_{0.975}s.e(\hat{\beta}_1)=0.7543\pm(1.96)(0.2704)=(0.2243,1.2843)$$

所以，Score 每增加一单位对应优势比的 95%置信区间是：

$$(e^{0.2243},e^{1.2843})=(1.2514,3.6121)$$

显然，变量 Score 的取值对股票价格是否产生涨幅有显著性的影响。

（3）模型预测

接下来将利用上表得到的 Logistic 回归方程进行股票价格涨幅预测，即利用研究窗口为 2015 年 3 月 1 号到 2015 年 6 月 9 号之间的数据进行模型预测检验。得到的检验结果如表 4.8 所示：

表 4.8 Logistic 回归检验结果

		y=0	y=1
中国银行	yhat=0	33	23
	yhat=1	5	16
农业银行	yhat=0	42	31
	yhat=1	1	12
交通银行	yhat=0	29	23
	yhat=1	8	18
招商银行	yhat=0	33	20
	yhat=1	11	18
浦发银行	yhat=0	38	25
	yhat=1	1	13

根据表 4.8 可以得到中国银行、农业银行、交通银行、招商银行与浦发银行的模型预测总体准确率分别为：63.6%、62.8%、60.3%、62.2%和 66.2%。总的来说预测的准确率还算可以，同时我们还可以在表 4.8 中看到对每家银行都存在一点规律：将不发生涨幅($y = 0$)而错误地预测为发生涨幅($\hat{y} = 1$)的天数比较少。

二、自相关 Logistic 回归分析

在上一节得到的模型中只考虑了协变量以及响应变量都不存在滞后的情况，本节将讨论响应变量本身有自相关或者新闻情感得分存在滞后的情况，即前一交易日股票价格波动情况可能会对当天交易股票价格波动趋势产生影响；或者新闻本身对于股市波动影响也存在滞后性。这些都需要在本节建模过程中进行探讨并得到最适合的模型。处理这类情况可以运用第三章第四节中提到的广义自回归模型，即自相关 Logistic 回归模型。

（1）模型估计

由于直观上并不能判断出会对现在的观测值产生影响的响应变量滞后项为几阶，本节将首先通过卡方检验的方法对不同滞后情况进行相关性检验，以此得到相对更显著的滞后项加入模型中。考虑到时效性问题，本文将要研究一至七阶的情况，按照中农交招浦的顺序得到不同银行最显著的检验结果见图 4.8：

```
> chisq.test(lag3)

        Pearson's Chi-squared test with Yates

data:  lag3
X-squared = 0.6538, df = 1, p-value = 0.4188
> chisq.test(lag6)

        Pearson's Chi-squared test with Yates

data:  lag6
X-squared = 6.1562, df = 1, p-value = 0.0131
> chisq.test(lag3)

        Pearson's Chi-squared test with Yates

data:  lag3
X-squared = 2.5249, df = 1, p-value = 0.1121
> chisq.test(lag1)

        Pearson's Chi-squared test with Yates

data:  lag1
X-squared = 0.77715, df = 1, p-value = 0.378
> chisq.test(lag4)

        Pearson's Chi-squared test with Yates
        correction

data:  lag4
X-squared = 0.55243, df = 1, p-value = 0.4573
```

图 4.8 各家银行相关性检验结果

根据图 4.8 中得到的相关性检验结果依次对每家银行进行自相关回归建模，在本文讨论的六个解释变量同时存在的情况下得到的自相关 Logistic 回归分析模型为：根据上一步基本回归分析的结果可以预料并不是六个解释变量都适合在模型中出现，对不同的银行剔除个别比较不显著的变量后得到五家银行新闻对其股票价格波动趋势影响的统计分析结果如下表 4.9 所示。其中本文采用了似然比检验来保证模型的显著性，似然比检验结果见于表 4.9 最后一列

表 4.9 自相关 Logistic 回归分析结果

	截距	Score	lnbank	lnsz	汇率	滞后项	LRT
中	81.898 (0.014)	0.849 (0.002)	16.829 (0.001)	-16.230 (0.001)	-10.075 (0.034)	-0.245 (0.107)	3.124 (0.077)
农	82.989 (0.011)	0.458 (0.099)	13.693 (0.004)	-11.270 (0.016)	-13.301 (0.006)	-0.418 (0.002)	7.687 (0.006)
交	84.237 (0.011)	0.936 (0.002)	16.307 (0.001)	-16.688 (0.000)	-9.318 (0.051)	-0.368 (0.007)	6.634 (0.010)
招	-2.697 (0.066)	1.716 (7e-05)	-0.422re (0.146)	0.986SHI (0.043)	*	-0.252 (0.078)	3.056 (0.080)
浦	57.808 (0.056)	1.309 (0.001)	7.411 (0.108)	-7.424 (0.099)	-7.629 (0.076)	-0.324 (0.021)	3.617 (0.057)

注：招行对应模型结果中出现 re 与 SHI 的表格分别代表变量热度与 SHIBOR 的估计值（并不是银行或者上证指数的估计值），*号代表该解释变量在此银行对应的模型中被剔除了，括号上面的数是 Logistic 回归系数的参数估计值，括号里面的数为回归系数 t 检验对应的 p 值。滞后项对应的滞后阶数与图 4.8 中的结果一一对应。

从表 4.9 可以看出，几乎每家银行对应模型的所有解释变量 t 检验都在在 10%的水平下全部都显著；且大部分解释变量与模型整体在置信水平 0.05 下都是显著的。其中本文关心的变量 Score 在除农业银行以外对应模型的 t 检验在置信水平 0.05 下都是显著的。且本节需要讨论变量 Score 是否存在滞后的情况，因此以农业银行为例进一步讨论该变量滞后一阶的情况，即前一天的新闻会对当天的股票市场产生影响。得到的模型结果如图 4.9 所示：

```
GLARMA Coefficients:
      Estimate Std.Error z-ratio Pr(>|z|)
phi_6  -0.4230    0.1303  -3.246  0.00117 **

Linear Model Coefficients:
          Estimate Std.Error z-ratio Pr(>|z|)
Intercept  85.0130   32.4398   2.621  0.00878 **
score       0.1173    0.2630   0.446  0.65560
汇率       -13.5952    4.8034  -2.830  0.00465 **
lnba        13.3600    4.7320   2.823  0.00475 **
lnsz       -11.0120    4.6068  -2.390  0.01683 *
```

图 4.9 农业银行 Score 变量滞后时自回归结果

从图 4.9 可以看出，变量 Score 滞后一阶后模型拟合的效果比原始模型中该

变量不滞后下的效果差，因此可以认为对本文研究的五家银行而言新闻本身对于股市波动影响可能并不存在滞后性。

（2）模型解释

根据表 4.9，以交通银行为例最终得到的自相关 Logistic 回归分析模型为：

$$\ln\left(\frac{P(\mathrm{Y}_t=1)}{P(\mathrm{Y}_t=0)}\right)$$

$$=84.237+0.936sc_t+16.307\mathrm{lnba}_t-16.688\mathrm{lnsz}_t-9.318\text{汇率}_t-0.368Y_{t-3} \tag{4.3}$$

由模型式可知，若情感得分 score 增加 1 分，股票价格产生涨幅的优势比为：$\exp(\hat{\gamma}_1)=\exp(0.936)=2.550$，即情感得分 Score 增加 1 分，股票价格产生涨幅的优势增加 100.550%。若三天前的股票价格产生涨幅，则当日股票价格产生涨幅的优势比为：$\exp(\hat{\beta}_1)=\exp(-0.368)=0.692$，即三天前的股票价格产生涨幅，则当日股票价格产生涨幅的优势降低了 30.8%。

其中该模型对应的参数估计结果显示如图 4.10 所示：

```
GLARMA Coefficients:
      Estimate Std.Error z-ratio Pr(>|z|)
phi_3  -0.3675    0.1369  -2.684  0.00726 **

Linear Model Coefficients:
          Estimate Std.Error z-ratio Pr(>|z|)
Intercept  84.2373   33.0738   2.547 0.010867 *
score       0.9361    0.3029   3.090 0.001999 **
汇率       -9.3176    4.7891  -1.946 0.051705 .
lnba       16.3070    4.8001   3.397 0.000681 ***
lnsz      -16.6878    4.7433  -3.518 0.000435 ***
```

图 4.10 交通银行自相关回归估计结果

因此对参数γ_1 (Score) 95%的置信区间是：

$$\hat{\gamma}_1\pm z_{0.975}s.e(\hat{\gamma}_1)=0.9361\pm(1.96)(0.3029)=(0.3424,1.5298)$$

所以，Score 每增加一单位对应优势比的 95%置信区间是：

$$(e^{0.3424},e^{1.5298})=(1.4083,4.6173)$$

显然，变量 Score 的取值对股票价格是否产生涨幅有显著性的影响。

（3）模型检验

接下来将利用上表得到的自回归 Logistic 回归方程分别进行股票价格涨幅预测，即利用研究窗口为 2015 年 3 月 1 号到 2015 年 6 月 9 号之间的数据进行模型预测检验。得到的检验结果如表 4.10 所示：

表 4.10 Logistic 回归检验结果

		y=0	y=1
中国银行	yhat=0	34	21
	yhat=1	5	17
农业银行	yhat=0	39	30
	yhat=1	7	10
交通银行	yhat=0	31	21
	yhat=1	9	17
招商银行	yhat=0	35	18
	yhat=1	10	19
浦发银行	yhat=0	37	21
	yhat=1	3	16

根据表 4.10 可以得到中国银行、农业银行、交通银行、招商银行与浦发银行的模型预测总体准确率分别为：66.2%、57.0%、61.5%、65.9%和 68.8%。与表 4.8 的准确率相比较发现除了农业银行的准确率稍有下降外，其余银行的准确率都有一点提高。同时对每家银行仍然都存在同一点规律，即将不发生涨幅($y = 0$)而错误地预测为发生涨幅($\hat{y} = 1$)的天数比较少。因此可以认为自相关 Logistic 回归模型比简单 Logistic 回归模型稍好一些。

第五章 结论与展望

互联网对人们的生活各个方面都有很大影响，其中一方面就是接收新闻消息的渠道变化。本文旨在利用文本挖掘相关技术、情感分析以及多项 Logistic 回归分析对互联网财经新闻可能会对股票市场价格波动趋势产生的影响进行一定的研究并将得到的模型用于预测。再研究过程中主要利用了爬虫技术得到文章研究所需的财经新闻文本数据，并通过情感分析方法对文本数据进行量化从而建立模型。本章将就之前的研究进行一个总体回顾，并对股票市场投资者和分析人员提出一些建议，同时也将就本文的不足做出一定的展望。

第一节　结论

本文通过情感分析以及回归分析的相关方法针对银行业五家上市公司在一段时间内发布于新浪财经上的个股新闻与其股票市场的价格波动趋势进行了一定的研究，得到了以下结论：

1、通过研究五家银行的情感得分、隔夜拆借利率、即期汇率、上证指数以及银行行业指数等解释变量对相应股票价格波动趋势进行 Logistic 回归分析以及预测检验可以得到一些结论，具体为：

（1）针对本文所研究的五家银行，再加入其它理论上会对股票价格波动产生影响的解释变量后，模型最终的结果都显示新浪财经上发布的个股新闻对应的情感得分对该日股票价格是否产生涨幅有显著影响。这首先说明本文采用的用于研究互联网财经新闻的情感分析方法有一定的可用价值；再者，该结论也说明了中国股民的投资行为会受到网络财经新闻的影响。其中，五家银行对应的模型中情感得分的参数估计值都大于 0，这说明新闻所表达的情感倾向越积极，对应的情感得分越高，则相应的该日股票价格发生涨幅的概率越大，这一结论与现实上的想法是一致的。

（2）从模型的预测检验结果中可以发现，通过 Logistic 回归与自相关 Logistic 回归得到的预测结果更多情况下都会认为股票不会发生涨幅。这一预测结果可以认为是该模型得到的结果比较保守，即相对于积极的新闻而言，消极的新闻

会对股票市场产生更大的影响。这一结论可以被理解为消极的、不利于公司的新闻会比较容易直接影响股票价格的波动，而当面对积极的新闻时，投资者更多时候可能会持一种暂时观望的态度，不会立即采取行动。

2、由于本文主要考察的是每日发布于新浪财经上的个股新闻对于股票价格波动趋势的影响，因此这也决定了可以放入模型进行研究的解释变量需要是以“天”为单位进行变化的时序数据。故本文最初考虑的解释变量个数有限，最终通过 Logistic 回归模型的实证研究过程从 6 个变量中最后都至少选出了包括新闻情感得分这一重点关注的协变量在内的两个协变量放入到模型中。利用 t 检验方法得到的结果显示，五家银行情感得分的系数估计值都为正数，与股票价格产生涨幅呈正相关关系。其中情感得分反映企业发布于财经网站上的新闻所代表的积极或者消极的程度：情感得分的增加通常说明发生了一些对企业有一定有利消息的举措或者政策等，理论上积极的消息会为股民带来对企业股票正面的看法，从而产生购买股票的行为；银行行业指数与上证指数在模型中大多很显著，这是可以预料到的，几乎任何一支股票的变动都不会脱离大盘与行业这两个整体指数的变动；而通常情况下，如果一国的货币升值，则股价便会上涨。汇率上升代表本币贬值，则汇率与股票价格发生涨幅成反比，这与模型中汇率的系数估计值为负数是一致的；研究本文得到的自相关 Logistic 回归模型可以发现，响应变量本身滞后项在模型中的估计值都为负数，因此可以认为若与当日响应变量最相关的滞后阶日股价发生涨幅，则当日发生涨幅的概率将下降。研究结果表明 Logistic 回归模型整体上以及单个变量对股票价格涨幅的影响都是显著的，同时变量的系数估计值的正负也比较符合经济或者现实理论，因此国内的股票市场投资者和分析者应该关心出现在财经网站上的新闻情感倾向。

第二节　建议

随着中国网络的不断发展与完善，互联网在人们的生活中正发挥着越来越多的作用。其中很重要的一方面正是通过财经新闻而对股市产生着影响。在本文第四章实证分析的基础上，本节将针对网络财经新闻对于股票市场价格波动趋势的影响为投资者、企业管理者和政府提出以下建议：

（1）对于股市投资者来说，在了解一个上市公司的运营情况和财务状况等公司内部情况基础上，需要时刻关注着公司的相关新闻。尤其对于那些报道的

可能会对公司带来消极影响的新闻需要格外关注，因为大多数情况下，消极新闻带来的股票价格下跌的可能性比积极新闻带来股票价格上涨的概率更高。

（2）对于企业管理者来说，则需要了解到在如今这个网络如此发达的时代，不应该只关注企业内部的管理。还要学习如何通过互联网的手段为企业股票市场创造更好的外部环境，通过这一渠道更好地向企业已有的或是潜在的投资者提供积极的信号，增强投资者的信心。

（3）对于政府来说，则更多的注意力应该放在监管上。通过本文的研究可以发现网络财经新闻会对股票市场价格波动趋势产生显著的影响，那么政府就需要保证发布于财经网站上新闻的准确性与时效性。既不能让一些别有用心的人利用这一手段欺骗投资者，又要为投资者提供透明的信息平台。

第三节　不足与展望

一、中文文本挖掘

本文在搜集到文本数据之后最重要的部分就是利用文本挖掘技术进行分词以及情感词典等方法进行情感分析。在这两个过程中，分词词典与情感词典的完善程度直接影响研究结果。而鉴于中文文本挖掘与情感分析的难度以及其导致的发展缓慢，本文在这两方面的研究分析还存在不足。

因此希望在以后的研究中，可以在中文分词与情感词典上做更深一步的完善，以减少这些因素可能给后期建模过程带来的影响。

二、样本的研究范围

本文研究的样本是几大银行在研究窗口内发生在新浪财经上的个股新闻标题，选择新闻标题而非正文内容时考虑的是正文中会出现更多关于非研究主体的文本信息，从而造成对研究主体的研究产生偏差。但是，不可否认这一选择也会在一定程度上浪费正文中的有价值的文本信息。

因此未来可以考虑在以后的研究中单独研究新闻正文内容，并思考如何将关于非研究主体的相关文本信息尽量减少，以降低对于研究主体的影响。

三、数据的处理

本文在得到每条新闻标题的情感得分后对数据进行了一定的预处理，其中包括将关于一个研究主体同一天内的新闻情感得分进行加总等。这部分的数据处理更多的是基于本人对相关研究内容的主观理解，并没有很多的文献借鉴或者理论依靠。

故未来在数据处理方面可以与相关领域内的专业人士进行研究讨论，获得更加专业性的处理方式。

参考文献

[1]陈华，梁循. 互联网股票新闻归类和板块分析的方法. 电脑开发与利用，2006 年第 11 期.
[2]陈伟. 基于时序文本挖掘的新闻内容理解与推荐技术研究. 浙江大学博士论文，2010:10-16, 26-35, 43-54, 76-88.
[3]韩春，田大钢. 对股票市场信息的文本挖掘. 中国高新技术企业，2008 年第 23 期.
[4]何诚颖. 中国股市市盈率分布特征及国际比较研究. 经济研究，2003 年第 9 期.
[5]胡凌云，胡桂兰等. 基于 Web 的新闻文本分类技术的研究，安徽大学学报：自然科学版，2010 年第 34 卷第 6 期.
[6]李国臣. 文本分类中基于对数似然比测试的特征词选择方法. 中文信息学报，1999, 14(3): 16-21.
[7]蔺璜，郭姝慧. 程度副词的特点范围与分类. 山西大学学报（哲学社会科学版），2003, 26(2): 71-74.
[8]孙春华. 情感表达对在线评论有用心感知的影响研究. 合肥工业大学博士论文，2012:1-10, 23-35, 43-44, 56-68.
[9]王素格，杨安娜，李德玉. 基于汉语情感词表的句子情感倾向分类研究. 计算机工程与应用，2009, 45(24): 153-155.
[10]杨继东.媒体影响了投资者行为吗？——基于文献的一个思考[J].金融研究，2007 第 11 期.
[11]杨娟. 互联网财经新闻对股票影响的实证分析——基于公司新闻语义分析的视角. 西南财经大学硕士论文，2012: 1-14, 25-32, 63-70.
[12]饶育蕾，王攀. 媒体关注度对新股表现的影响——来自中国股票市场的证据[J]. 财务与金融，2010 年第 03 期.
[13]饶育蕾，彭叠峰，成大超. 媒体注意力会引起股票的异常收益吗？——来自中国股票市场的经营证据，系统工程理论与实践，2010 年第 30 卷第 2 期.
[14]赵丽丽，赵茜倩，杨娟，李庆. 财经新闻对中国股市影响的定量分析. 第七届全国信息检索学术会议(CCIR2011)，2011 年 10 月.
[15]赵丽丽. 互联网财经新闻对股市影响的定量分析. 西南财经大学硕士论文，2012: 1-10, 21-22, 59-60.
[16]赵伟，梁循. 互联网金融信息量与收益率波动关联研究.计算机技术与发展，2009 年第 19 卷第 12 期..
[17]Beaudry, P. and Portier, F. Stock Prices, News and Economic Fluctuations. 2004.
[18]Fama, E.F. The Behavior of Stock-Market Prices. Journal of Business, vol.38, no.1, pp.34-105, 1965..
[19]Gunn, S.R. Support Vector Machines for Classification and Regression. Citeseer, 1998.

[20]Ludwig Fahrmeir, Gerhard Tutz. Multivariate Statistical Modelling Based on Generalized Linear Models. Published by Springer-Verlag New York, Inc, 2001.

[21]Manolis Maragoudakis, Dimitrios Serpanos: Exploiting Financial News and Social Media Opinions for Stock Market Analysis using MCMC Bayesian Inference [J]. J Computational Economics, pp.1-34, 2015.

[22]Niederhoffer, V. The Analysis of World Events and Stock Prices', The Journal of Finance, vol. 44, no.2, pp.193-219, 1971.

[23]Perold, A.F. The Capital Asset Pricing Model. The Journal of Economic Perspectives, vol. 18, no.3, pp.3-24, 2004.

[24]Salton, G. and Yu, C.T. On the Construction of Effective Vocabularies for Information Retrieval, ACM SIGIR Forum, vol.9, no.3, pp.48-60,1974.

[25]Schumaker, R.P. and Chen, H. Textual Analysis of Stock Market Prediction Using Breaking Fianacial News-The AZFinText System. ACM Transactions on Information Systems (TOIS), vol.27, no.2, pp.1-19, 2009.

[26]Tan, A. H., Text Mining: The State of the Art and the Challenges. Proceedings of the PAKDD 1999 Workshop on Knowledge Discovery from Advanced Databases, pp.65-70, 1999.

[27]Tetlock, Paul C. Giving Content to Investor Sentiment: The Role of Media in the Stock Market. Journal of Finance, pp. 1139-1168, 2007.

[28]Tetlock, P., Saar-Tsechansky, M. Macskassy, S. More Than Words: Quantifying Language to Measure Firms' Fundamentals. Journal of Finance. VOL. LXIII, no.3, pp.1437-1467, 2008.

[29]Wiithrich, B., Permunetilleke, D., Leung, S., Cho, V., Zhang, J., and Lam, W., Daily Prediction of Major Stock Indices from Textual WWW Data. KDD-98 Proceedings, 1998.

[30]Xiqian Zhao, Juan Yang, Lili Zhao, Qing Li. The Impact of News on Stock Market: Quantifying the Content of Internet-based Financial News. IDSI2011, Taipei, 2011.

[31]Yang, Y., An Evaluation of Statistical Approaches to Text Categorization. Journal of Information Retrieval, vol.1, oo.69-90, 1999.