上海财经大学

毕业论文

题目 文本挖掘技术在股评信息情感判断中的应用

姓　　名　孔潇

学　　号　2014210714

系　　别　统计与管理学院

专　　业　应用统计硕士

定稿日期　　年　　月

文本挖掘技术在股评信息情感判断中的应用

摘要

文本挖掘是数据挖掘的一个重要分支，而在金融市场，对金融文本信息的挖掘和其商业价值的探索正在处于快速发展的阶段。而对金融文本信息的情感判断则是整个文本挖掘流程及量化策略构建中的核心环节。金融文本信息主要包括新闻、股评、研究报告三大类别，相对于其他两种文本信息，股评信息的获取、处理和分析的难度最大，但其信息量也最大、也最能直接反映投资者的情绪，无论从理论上还是从投资实践上都更具有研究意义。

本文借助于申万宏源的大数据舆情系统，使用东方财富股吧的股评信息进行情感判断的研究。本文通过文本挖掘技术，包括分词、分句手段将文本信息切割成单词后，通过两方面的尝试试图找到最适合中国金融市场股评信息情绪判断的方法。一方面尝试不同的文本特征选择方式，包括词频法、互信息法、卡方方法；另一方面比较不同的情感分类算法，包括情感词典法、逻辑回归、朴素贝叶斯、支持向量机、随机森林。结果发现，从文本特征选择的角度看，卡方方法要明显优于其他两种特征选择方法；，而在情感分类算法上，朴素贝叶斯、逻辑回归、支持向量机方法的正确率达 70%以上可以满足商业要求，而支持向量机的分类准确性又相对最优，达到 75%的分类准确性。

关键词：金融文本挖掘 情感判断 特征选择 朴素贝叶斯 支持向量机

Abstract

Text mining is an important branch of data mining. it is in a stage of rapid development for the mining of financial text information and the exploration of its commercial value. The emotional judgment of financial text information is the key link of text mining process and quantitative strategy construction. Financial text information mainly includes three categories: news, stock messages, research report. Stock messages most directly reflect the investor sentiment, whether in theory or from the investment practice is more significant than other two text information.

With the help of Shenwanhongyuan public opinion big data system, using Eastern wealth stock comment information to conduct emotion judgment research. The author attempt to find the most suitable emotion judgment method for China's financial market. Firstly, to try different text feature selection method; Secondly, compare different sentiment classification algorithm. It is found that the chi square methods is significantly better than the other two feature selection methods; naive Bayes, logistic regression, support vector machine has correct rate of more than 70% can meet the requirements of commercial, and support vector machine classification accuracy and relatively optimal, reached 75% of the classification accuracy.

Key words: financial text mining; feature selection; emotion judgement; Naïve Bayes; Support vector machine

目 录

第一章 序言

第一节 选题背景

文本挖掘是数据挖掘的一个重要部分，其研究对象不同于以往的结构化的数值型数据，而是非结构化的文本型数据，在国内主要集中在互联网中的微博、新闻、用户评价等相关数据的研究上上。在金融领域，金融文本挖掘正日益吸引着越来越多人的关注，其所蕴含的潜能和商业价值正被日益发掘出来，成为非常热门的领域，原因有以下几点：

一是对传统数值型数据的研究已经相对成熟，从基本面研究人员对公司的财务数据、行业研究人员对公司所处行业的行业数据、和公司运行相关的宏观经济变量数据，到技术分析者运用的股票量价信息以及衍生而来的各种数值型技术指标，都被较充分地应用于对股票市场的分析中；而对金融文本信息的挖掘和使用还处于刚刚起步的阶段，尤其在中国市场上，几乎没有投资者真正在使用文本信息进行投资，是非常有潜力的一个领域；

二是从网络的文本信息中，能够直接了解到投资者的投资意见。传统的金融领域投资者，通过股价的变动、量价的律动、图表的形态来判断投资者对待市场的态度，比如在量价齐涨测阶段，认为投资者具有很大的投资热情。可以看到，这种判断是间接的、后知后觉的、揣测性质的，不能精确、直接地看出投资者的投资意见。而通过文本挖掘的方式，通过股市人员的研究报告可以获取研究员对股市的态度和意见；通过股吧里投资者的发言可以获知散户对市场的态度；通过新闻的内容和大家在互联网上的讨论可以找到投资者主要关注的主题、热点等等。这种通过文本信息对投资者意见的研究，是直接、可测可量化、实时的，相对于传统的研究方式有明显的优势；

三是由于国内互联网的发展日益成熟，为金融文本挖掘提供了充足的研究数据。一方面，由于互联网的普及，越来越多的散户会通过网络世界来表达自己对市场的看法，接触股市相关的新闻，在网络世界里留下越来越多的文本信息；另一方面，网络也成为股票市场和投资者进行沟通的重要渠道，包括券商的研究报

告、上市公司的公告、媒体人关注的公司新闻都会在网络中留下文本的信息。迄今为止，网络金融文本无论从数量上还是时间跨度上，都能够充分满足金融量化的研究需求。

第二节 研究意义

国内对金融文本数据的应用还处于刚起步的阶段，在诸多方面都还很面临着较多研究挑战。首先是数据获取方面，金融文本数据包括微博、新闻、股吧、研究报告等等网络文本信息，散布在互联网的各个角落，几乎没有办法获取互联网的全部信息，如何能够尽可能获取充足的、具有代表性的金融文本数据就是学界和业界共同面临的一个复杂技术难题。其次在对金融文本信息的分析上面，国内进行的研究也寥寥无几，无论是金融文本信息的情感判断还是市场投资者观点和股票市场的关系的研究都处于非常初始的状态。研究尚且如此，更不用说对这些金融文本信息的商业价值的应用。

而在这些诸多的研究课题中，对金融文本信息的情感判断是整个文本挖掘及量化投资模型构建过程中及其重要的一个环节。投资者对股票的评价是好是坏、市场整体对该股票的评价是正面为主还是负面为主、近期对该股票的评价有没有发生积极的变化从而可能存在相应的交易机会，这些都离不开对股票评论、对股票新闻准确的情感判断。其本身具备相当有潜力的商业价值。对股民、对研究员的投资态度的准确把握，会使得主观投资者对股票市场阶段的判断有着更坚实的基础；也给量化投资者提供了新的测量信号，为构建量化投资策略打开了全新的视角。

而众多的股票信息类型中，股评信息是直接反应中小股民投资情绪的信息类型。其他类型的股票信息，如公告、新闻，本身可能具备积极或者消极的信号，但是由于股票市场的复杂性，看似好的新闻或者公告最后作用到股民身上并不一定能激发股民相对应的情绪，是数据相对间接的网络文本信息类型。譬如说股票市场中的公司公告信息，当公司发布了一条公司盈利同比增长 20%的公告，则从文本挖掘的处理方法上，无论如何这都会是一条看好的信息，但是如果市场上绝大部分投资者之前的预期是 50%的增长率，那么这条信息最后反映在投资者的投

资态度上其实是负向的。所以本文的研究对象将选用股评信息，因为其是股民心态最直接的体现。通过研究此信息的情感倾向，能够正确对中小股民的投资情绪或投资意见进行识别和描述，对进一步的投资研究有着非常重要的实践价值。

第三节 文献综述

在金融文本挖掘这一领域，国外研究人员已经积累起一定的研究成果，而国内商处于刚刚起步的研究状态，接下来将主要以国内外研究的发展历程来介绍前人的相关研究文献。

在国外，基本上在网络文本信息对金融市场的影响研究中，遵循数据处理难度的由易到难、数据形式结构化到非结构化大致经历了以下阶段：

在早期的金融文本挖掘研究中，一般研究人员使用的研究工具基本为网络信息中自带的数量化信息，比如发帖量、跟帖量、点赞数或者点踩量等可以直接获取的数值型数据，这种数据获取容易、分析简单，不涉及对具体文本内容的深入研究。

Wysocki（1998）对美国股市中，互联网上股市相关的发帖量和股票市场的关系进行了开创性的研究，其从截面和时间序列两个维度对这种影响是否存在进行考察，验证发帖量的变动和公司基本面及股票市场活动情况是否相关。通过对 Yahoo!message 论坛上的 3000 只股票的信息进行研究，发现过去股票累积收益率高、财务状况好的公司其论坛发帖量相对较高；而日发帖量的变化则与日股票收益率和成交量成明显的正相关关系，且前日发帖量可以用来对第二天股价和成交量进行一定程度的预测。

Tumarkin 和 Whitelaw（2001）则进一步考察了网络股票论坛活跃度和股票超额收益和成交量的关系。他们使用了 RagingBull 网站的股票讨论区的数据作为研究对象，这个网站的股票讨论区允许股民对投资者的意见进行打分，从而利用这些分数和发帖量能够度量投资者的态度以及活跃度。在研究中他们发现，当对某只股票的评论突然开始活跃了起来，并且对其的评价有一个明显的提升，那么该只股票往往会伴随着产生一定的超额收益。但是这种伴随的关系却无法转化成能够预测的模式，Tumarkin 和 Whitelaw 在经过研究后发现，很难通过这种活跃度

的变化来预测超额收益或者是成交量变化，从而得出市场有效性的结论。

Da 和 Engelberg（2011）年更是使用了谷歌的搜索数据对股票的关注度进行度量，并且研究其关注强度和股票量价的关系。其创新性地使用谷歌里面关于股票的搜索频次，建立了一种新的市场关注度指标——Search Volume Index(SVI)。在对罗素指数中的 3000 只股票进行了研究之后，发现 SVI 和其他市场上已有的、通过股票量价等方式定义的股票关注度指标具有一定的相关性却又不完全相同，相对来说 SVI 相对于其他描述股票关注的关注度指标都更具备实时性；而且 SVI 在某些情况可以作为股价的先行指标，譬如在 SVI 上升之后的两周内，对应的股票往往能走出更高的价格，而在一年之内这种价格的上升又最终得以回复。

Joseph（2011）同样使用了谷歌的搜索数据，利用对股票单词的搜索频次来代表投资者的关注强度，不过其检验股票关注强度和股票未来收益方法是采用金融领域的排序打分方法。具体做法，即在每一期期末，根据股票的关注强度，将股票等分为 5 个等分，再比较下一期五个股票组合的投资收益情况。结果发现，关注强度大的股票组合收益率相对来说会更大，验证了关注强度大股票组合存在超额收益。

随着计算机技术的不断发展和文本挖掘技术的日臻成熟，越来越多的研究者开始脱离原来数量化的研究手段，转而关注金融文本的内容，而试图分析金融文本的内容信息、利用投资者的投资情绪，并综合金融文本的数量化信息共同研究对股票市场的影响。

Antweiler 和 Frank(2004)即利用股票评论的情感信息研究其与股票市场量价的关系。他们使用道琼斯工业指数和道琼斯互联网公司指数成分股的共 45 家公司进行研究，而数据的来源是 2000 年 Yahoo 金融和 Raging Bull 的论坛数据。为了对大量的股评进行情感判断，作者手动提取了股评中的 1000 条信息作为训练集，将者 1000 条信息分为看多、看空、中立三种态度，接着综合使用了朴素贝叶斯、支持向量机两种算法对所有 160 万条股评信息进行分类，从而得到所有信息的情感方向。但是作者并没有对这种分类方法进行效果的检验，没有用检验集测试实际分类成功率，而对文本情感判断这个步骤是整个文本挖掘流程中至关重要的一环。在得到股评信息的情感方向后，Antweiler 和 Frank 综合得到了每只股票的每一天的情感值、认同的一致性水平、股评量，对以下三个问题进行了实证

研究：股评信息可以用于预测收益率吗？股评信息的一致性水平可以影响股票成交量吗？股评信息可以用于预测波动率水平吗？而通过研究发现，通过股评信息很难预测股票的收益率；而越高的投资者分歧（即对股票情感的不一致水平）则确实伴随着更高的成交量；并且股评信息确实能够帮助对股票的波动率进行预测。

除了对网络论坛等的研究，还有学者就金融新闻进行文本挖掘并进行金融相关的研究。Paul（2005）为了量化研究新闻媒体对股票市场的影响，对华尔街日报的新闻专栏进行了相关研究。他首先对华尔街日报专栏的新闻进行统计，自己设计产生每日的新闻情绪指数，并使用向量自回归（VARs）对新闻情绪指数和股票市场活动的关系进行探索。结果发现，非常乐观的新闻情绪能够稳定地预测接下来的股票指数下行压力；并且一般水平的乐观或者二悲观新闻情绪能够预测市场接下来较高的成交量；最后，当市场的收益率变低时，往往伴随的是之后的高涨的新闻看多情绪。

国外学者也有人专门就金融新闻信息的情感判断做过深入研究。Bozic 和 Seese（2011）通过构建神经网络分类器对金融新闻信息的多空情绪进行研究和判断。利用分类结果，他们发现金融新闻的情感水平和股票未来的收益率在统计上有显著的关联。

随着社交平台的快速发展，也有研究人员开始使用非金融论坛的文本信息进行股票研究。Bollen、Mao 和 Zeng（2010）使用 Twitter 上面的普通用户微博对整个 Twitter 使用者的情绪程度进行判断，注意这里，他们使用的已经不是股票相关的评论而是所有用户的信息，判断的也是所有用户的整体的情绪状态。作者使用了 2008 年间 Twitter 用户的评论，并借助于谷歌的两套情感分类系统自动地对这些情绪进行分类。其一是谷歌的 OpinionFinder 情感分类器，可以将文本分为积极和消极两类情感；其二是 Google-Profile of Mood States(GPOMS)，可以将文本在冷静、警惕、确定、生动、善良、高兴六个维度上进行量化度量。再经过这种详细的文本情感处理之后，和股票市场的具体表现进行研究。结果发现，在某些情绪维度上，道琼斯工业指数的涨跌能够较好的被预测，而在另外一些维度上面则没有效果。其最终对道琼斯指数的涨跌的预测能够达到 87.6%的准确度。这里面还值得一提的是作者对 Twitter 文本的细致处理。譬如其在对文本进行情感判断时，并不是使用全部的 Twitter 微博，而是选取了只包含能明确表达情感

的词组，如‘I feel’，‘I am feeling’，‘I don’t feel’，‘im’，‘makes me’。随着互联网社交媒体的不断发展，文本处理日益面临的问题不再是数据量不够的问题，而是数据量太大如何保证正确性的问题，通过以上的这种做法虽然会损失大量数据，但是却能保证分析的准确有效，是互联网文本挖掘处理的一个原则的体现，即牺牲一部分数据量以保证分析的准确性。

还有学者在文本挖掘的特征提取上面使用更加细致的方法以提高文本分类的正确性。Ormos 和 Vazsonyi（2011）分析了 10 年间股票回报率和股票新闻之间的关系，所使用的数据源是 LexixNexis 数据库中当天的七家金融杂志的新闻公告。其处理文本的方法较为不同，与一般采用单词的方式不同，Ormos 和 Vazsonyi 运用语法分析的方法，只使用了具备形容词—名词结构的词组，作为用来对新闻公告的情感进行判断，分类为正向、负向两类情感，并且比较了一般方法，确实取得了更好的效果。之后利用处理好的新闻情感信息，作者只选取历史上出现极端涨跌的情况下，结果发现在这种极端行情下，利用新闻公告得到的情感值能够较好且持续性地对次日的收益率进行预测。

国内，网络文本数据对股票市场的影响的研究刚刚起步。马俊伟（2014）利用东方财富股吧中公司板块的论坛文本信息作为研究对象研究股吧信息量代表的股民活跃度对股票市场的影响。通过时间序列上面相关的模型，得到以下结论：股吧中的网络文本量和次日的股票收益率之间有密切的关系，且与股票的波动率有明显的相关关系。

赵丽丽等（2012）则使用了新闻数据，研究新闻文本对中国股市的影响。其采用的文本数据是和讯网和新浪财经网的新闻文本数据，再通过 TF-IDF 方法将新闻文本转化为量化的向量，从而利用新闻文本数据和股票收益率数据进行 SVR 建模。实证结果表明，新闻文本对股票市场的影响力和持久度在不同市场有所不同，于深市股票的影响要大于沪市股票；且市值比较小的公司受到新闻的影响要更大。

祝宇（2013）同样使用了东方财富股吧的文本信息进行相关研究，不过其不是只利用了股评信息的数量，还关注了股评的内容，在对股评进行情感判断之后使用了网络股民的投资者情绪。其通过朴素贝叶斯方法对股评信息进行正向、负向、中立的态度判断，之后自己构建了投资者情绪指数，并利用投资者情绪指数

和股评文本量共同研究与股票量价的关系，结果发现：网络文本信息对股票市场的量价情况有一定程度的预测作用，发帖量越大，往往相关股票的波动率、成交量都会更大，且影响周期相对较长；而情绪指数对收益率的影响则相对较弱而且时效较短，并且情绪指数处于高位的时候股价的波动率也较高。

第四节 研究基本思路和方法

首先借助申万宏源股票舆情系统获取过去数月主要股票论坛的股票评论数据，从得到的股票评论数据中随机抽取 2000 条左右的股票信息作为建模样本，依次进行以下处理：

使用文本挖掘技术对初始语料进行处理，包括分句，即将段落划分成更易分析的短句;之后是分词,将中文短句进一步地切分成中文单词;最后是特征选取，本文将采用三种主流特征选取方法对文本特征进行筛选,选出信息量最大的部分单词，并于之后的检验中比较不同特征选取方式的效果。进行特征筛选之后，即将每份股评的信息向量化，便于下一步的数量化建模。

接下来是对金融文本的情感判断进行数据挖掘建模，将使用以下五种方法分别进行情感判断研究：1）证券行业当下主要使用的情感词典方法；2）传统的经典方法——朴素贝叶斯方法；3）逻辑回归模型；4）支持向量机（SVM）方法；5）随机森林方法，试图将股票信息所包含的情感进行分类。之后在抽取出来的验证样本股票信息中进行检验，比较各种方法的分类效果，试图找到适合中国股票市场股评文本信息情感判断的最佳方法。

第五节 重难点和创新

难点一方面在于数据的获得方面，股评信息散落在网络的各个论坛之上，这里如何从网络中将待研究的数据获取出来本身具有一定的难度;另一方面在于数据形式的特殊性，文本数据无论在数据存储、数据清洗、编程操作上都比传统的结构化数据要更加困难，对计算机编程操作有较高的要求。

在创新方面，本文选用股票网络信息的股评信息进行研究，更加直接对投资者情绪本身进行研究，而不是从新闻、公告等间接文本信息进行研究，更具实践价值；且抽样使用的数据来自国内主要论坛股吧的全部信息，非常具有代表性，结果相对来说更加符合规范。

另一方面，从研究层面上面来说，现今国内业界对股票网络信息的研究和应用都未深入到文意挖掘中，基本上都停留在股票信息的数量、点赞数等数量方面的信息，而未深入到股票信息的内容的研究上；而就方法论上面来说，当前业界基本上只用情感词典的方式进行判断研究，学界在国内迄今只尝试过朴素贝叶斯方法，笔者将在国内业界和学界的研究基础上，尝试使用并比较更多的文本挖掘方法，找到最适合中国股票市场的股吧信息情感判断方法。

第二章 模型介绍

第一节 文本处理技术

本文使用的文本处理技术，包括分词、分句和特征词提取，以下将详细介绍这三种文本处理技术。

一、分句

在获取一段中文语料之后，首先要进行的处理工作就是为该语料进行分句，将一个段落分成句子。这里使用的分句方法是，将一个段落看成是长字符串，借助于“,.!?:;~，。！？：；～”等标点符号作为分割标志，将段落分割成句子，具体算法如下：

a. 搜寻剩余段落中第一个不是标点符号的字符作为起始字符

b. 搜寻起始字符之后第一个标点符号的前一个字符作为终止字符

c. 将起始字符和终止字符及其间的所有字符取出作为一个新句子

d. 将终止字符前的所有字符删除并重复前面三步

二、分词

分词方法大致可以分为三类，基于词典匹配的分词方法；基于词频统计的分词方法；基于知识理解的分词方法。

本文主要使用的方法即是基于词典匹配的分词方法。该方法通过使用积累的词典、汉语知识等现有的单词库，通过匹配的方式对语料进行分词，基本原理是将语料的汉字字符串取出，在词典中去寻找一样的单词，找到即为匹配成功。如：正向匹配法、最大匹配法等。以下逐一介绍几种基于词典的匹配分词方法。

（1）逐词遍历法：依次将词库里面的单词单独提出来，在语料中对句子进行匹配，匹配的顺序按单词长短由大到小。这种分词方法只适合词库比较小的时候才能保证效率，因为无论如何都需要将词典全部遍历一遍。

（2）最大匹配法（正向）：首先确定词典中最长单词的长度 n,然后对语料的字符从左到右进行匹配，从左边第一个长度为 n 的字符开始，在词典中进行匹配，如果匹配成功，则将该字符串作为一个单词；如果匹配失败，则将该字符串最右边的字符去掉继续在词典中匹配，知道匹配成功为止

（3）逆向最大匹配法：和前面介绍的最大匹配法正好相反，在确定最长单词长度 n 之后，从文档某端开始，取最后面的 n 个字符在词典中进行匹配，如果匹配成功则继续，匹配失败去掉 n 个字符的第一个字符继续匹配直到匹配成功为止

（4）双向匹配法：双向匹配法即是对正向匹配法和逆向匹配法进行结合后的匹配方法，其先找出语料中的标点符号，再通过这些标点将语料切分成小段，再在这些小段中同时使用最大匹配法和最小匹配法。如果两种方法得到的结果一样，那就取该结果为匹配结果；如果结果不一致，则取最小集作为结果。

第二节 特征选择

选择合适的单词

在处理文本数据的时候，不可能把语料中的所有单词作为自变量进行建模，需要在所有的单词中进行选择，把信息量更多、更重要的单词保留下来，即实现空间维度的降维。文本挖掘中的特征选择方法有六种，DF, MI, IG, CHI，WLLR,WFO，以下将逐一介绍其中较为常用的四种特征选择方式。

首先进行用于进一步解释的概率定义：

$p(t)$:文档 x 中存在单词 t 的概率

$p(\overline{c_i})$:文档 x 不属于类别 c_i 的概率

$p(c_i \mid t)$：在文档 x 中存在单词 t 的前提下，该文档属于类别 c_i 的概率

$p(\bar{t} \mid c_i)$:在文档 x 属于 c_i 的前提下，文档中不包括单词 t 的概率

这里给出一份样本统计，说明以上概率的估计方式 $\overline{c_i}$：

表 2.1 特征选择概率说明表

类别 / 单词	c_i	$\overline{c_i}$	总数
t_j	A	B	A+B
$\overline{t_j}$	C	D	C+D
总数	A+C	B+D	N

其中，A：包括单词t_j且属于文档c_i类别的文档数目

B：包括单词t_j且不属于文档c_i类别的文档数目

C：不包括单词t_j而属于文档c_i类别的文档数目

D：不包括单词t_j且不属于文档c_i类别的文档数目

相应的概率计算为：

$$p(t_j) = (A + B) / N$$

$$p(c_i) = (A + C) / N \qquad (1)$$

$$p(c_i \mid t_j) = A / (A + B)$$

其他概率计算方法类似。再获取这些概率之后，下面结合这些概率介绍四种特征选择方法

一、DF（Document Frequency）

DF：包含某个单词的文档数目，一次来衡量该单词的信息量，DF 的公式如下：

$$DF = \sum A \qquad (2)$$

DF 方法的特征选择理念即是，如果这个词在文档中出现的次数比较多，那么它包含的信息量就比较多。而出现次数很少的单词，不论它和文档类别的关系有多大，对文档的分类影响也很小。但是 DF 方法的缺陷也非常明显，首先其属于无监督的学习算法，无法利用语料的类别信息；另一方面可能会选出一些无意义的，仅是出现频次比较多的单词。

二、MI（Mutual Information）

MI（互信息法）结合单词出现的概率和单词在某类别文档中出现的概率对重要性进行判断，其公式为：

$$MI = \log \frac{p(t_j \mid c_i)}{p(t_j)} \tag{3}$$

仅针对第i个文档的特征选取？

可以看到，如果一个单词出现的频次不大，而取偏偏在某类文档中出现的频次较高，那么用这种方法就认为该单词具有显著的分类效果。互信息法属于有监督的方法，会利用文档本身的类别信息；但是同时也有一个比较明显的问题，就是如果该单词出现的频次比较高，那么就有可能计算出来的信息量较小，而如果该单词携带的信息量比较高，就会判断失误

三、IG(Information Gain)

信息增益法，即是通过比较包含和不包含该单词的文档中，该单词提供的信息量，综合两部分信息得到对该单词的判断。具体公式如下：

$$G(t_j) = p(t_j)[\sum_{c_i} p(c_i \mid t_j) \log p(c_i \mid t_j)] + p(\overline{t_j})[\sum_{c_i} p(c_i \mid \overline{t_j}) \log p(c_i \mid \overline{t_j})] \tag{4}$$

含有tj词的类中的information gain

可以看到，在互信息法中，只考虑了存在该单词的文档的概率，而信息增益法还考虑了不存在该单词的文档概率。

这个貌似只是引入tj后的熵的公式吧，还要减去引入前的熵才对呀。但是如果是来选取各个G(tj)中最高的那些，公式中含有的引入前的熵是一样的

四、CHI(Chi-square)

方差选择法利用的是统计学中的卡方检验的方法，原假设即是，单词和文档的类别是不相关的。如果计算出来的卡方值越大，即认为这种相关性越明显，越有充分的理由认为单词在文档分类中有较大的作用，计算公式如下：

$$X(t_j, c_i) = \frac{N(AD - BC)^2}{(A + C)(B + D)(A + B)(C + D)} \tag{5}$$

第三节 情感判断方法

文本挖掘中的情感判断是非常重要的一类分析技术，大体上包括两大类方法，一类是传统的情感词典方法；一类是基于统计理论的数据挖掘方法。以下介绍本文使用并进行比较的五种情感判断方法。

一、情感词典方法

通过给定情感词典，将段落中的每一个词语的情感进行标示，如果该词语的情感是正向的，那么给予该词语+1 分，如果该词语的情感是负向的，给予该词语-1 分。通过为段落中的每一个词语标分，并且累加得分，从而得到整个段落的情感总分。如果情感总分大于 0，则判断该段落为正向情感，反之判断为负向情感。

使用情感词典方法的一大前提是有相关领域完备的情感词典。本文使用的情感词典，结合了知网情感词典和长江证券金融情感词典，力求覆盖网络股吧的用词。以下给出长江证券研究所金融工程范辛亭等（2014）使用的正负情感词典：

表 2.2 长江证券金融情感词典—利多词库

利多词库				
刺激	利好	爆发	关注	机会
转机	时机	有望	引爆	有利于
推荐	坚定	看好	催生	巨大
净流入	井喷	反弹	盛宴	冲刺
支撑	瞄准	超预期	大增	

表 2.3 长江证券金融情感词典—利空词库

利空词库				
地雷	不利	利空	失败	解禁
跌停	失利	围剿	看空	风险
谨慎	无望	恶意	亏损	套现

二、逻辑回归方法

一般的线性回归模型研究 E(y) 和 $\vec{x}$ 之间的线性关系，假设 $y=\vec{\beta}\vec{x}+\varepsilon$，其中 y 一般是连续变量且 $Y|\vec{x}\sim N\{u(\vec{x}),\sigma^2\}$，从而可根据 $u(\vec{x})=\vec{\beta}\vec{x}$，代入正态分布的概率密度函数从而得到似然函数，并通过最大化似然函数的方式来得到$\vec{\beta}$的估计。而当 Y 只取二值的时候，Y 不连续，因此很明显 Y 不会服从正态分布，在这种情况下来应用回归分析，需要进行适当地变形。假设 Y 服从伯努利分布，纳入$\vec{x}$的影响，认为 $Y|\vec{x}\sim bin\{1,p(\vec{x})\}$，于是可以在 $p(\vec{x})$ 和 $\vec{x}$ 的之间建立回归模型。模型需要对 $p(\vec{x})$ 的形式进行变化，因为 $p(\vec{x})$ 的取值在 0 和 1 之间，而 $\vec{\beta}\vec{x}$ 的取值却是不确定的，因此把 p 假设成 $\vec{x}$ 的线性函数或是多项式函数都是不合适的。对 $p(\vec{x})$ 进行 LOGIT 变换：

$$f(p(\vec{x}))=\ln\frac{p(\vec{x})}{1-p(\vec{x})} \tag{6}$$

其中 $\frac{p(\vec{x})}{1-p(\vec{x})}$ 是“事件发生”比“事件没有发生的”优势。通过 LOGIT 变换，$f(p(\vec{x}))$ 的取值在 0 和 1 之间，所以可以把 $f(p(\vec{x}))$ 设为 $\vec{x}$ 的回归函数，建立线性回归模型如下：

$$\ln\frac{p(\vec{x})}{1-p(\vec{x})}=\vec{\beta}(\vec{x}) \tag{7}$$

根据 $Y|\vec{x}\sim bin\{1,p(\vec{x})\}$，可以得到 Y 的概率密度函数，即 $p(\vec{x})^y(1-p(\vec{x}))^y$。将 LOGIT 变换后建立的 $p(\vec{x})$ 和 $\vec{x}$ 的关系式代入上式，就可以得到似然函数。对似然函数取对数即可以得到下式：

$$
\begin{aligned}
l(\vec{\beta}) &= \sum_{i=1}^{n}[Y_i \log\{p(\vec{x}_i)\} + (1-Y_i)\log\{1-p(\vec{x}_i)\}] \\
&= \sum_{i=1}^{n}[Y_i \log\{p(\vec{x}_i)/(1-p(\vec{x}_i))\} + \log\{1-p(\vec{x}_i)\}] \\
&= \sum_{i=1}^{n}[\vec{\beta}\vec{x}_i Y_i - \log\{1+\exp(\vec{\beta}\vec{x}_i)\}]
\end{aligned} \tag{8}
$$

通过最大化上式即可得到 $\vec{\beta}$ 的估计。

得到参数估计之后，进而可以计算新样本的判断概率 P。

三、朴素贝叶斯方法

$$
P(C \mid X) = \frac{P(X \mid C)P(C)}{P(X)} \tag{9}
$$

将表示成为向量的待分类文本 $X(x_1, x_2, \ldots, x_n)$ 归类到与其关联最紧密的类别 $C(C_1, C_2, \ldots, C_J)$ 中去。其中， $X(x_1, x_2, \ldots, x_n)$ 为待分类文本的特征向量，$C(C_1, C_2, \ldots, C_J)$ 为给定的类别体系。设属于给定类别 $C_1, C_2, \ldots, C_J$ 的概率为 $(P_1, P_2, \ldots, P_n)$。则 max$(P_1, P_2, \ldots, P_n)$ 所对应的类别就是文本 X 所属的类别，因此分类问题被描述为：求解方程（2）的最大值。

这里写错了吧

$$
P(c_j \mid x_1, x_2, \ldots x_n) = \frac{P(x_1, x_2, \ldots x_n \mid c_j)P(c_j)}{P(c_1, c_2, \ldots c_n)} \tag{10}
$$

由于分母值不变，所以哪个类别计算出来的分子值最大，就把文本划分到哪一类中。而根据朴素贝叶斯的假设，文本特征向量属性 $x_1, x_2, \ldots, x_n$ 独立，其联合分布等于各个属性特征概率分布的累积，即

$$
C_{best} = \arg\max_{c_j \in C} = P(c_j)\prod_i P(x_i \mid c_j) \tag{11}
$$

VC维理论？SVM是建立在VC维理论上的？

四、支持向量机

支持向量机方法是建立在统计学习理论的 VC 维理论和结构风险最小原理基础上的，根据有限的样本信息在模型的复杂性（即对特定训练样本的学习精度，Accuracy）和学习能力（即无错误地识别任意样本的能力）之间寻求最佳折衷，以期获得最好的推广能力。

设分类函数的表达形式为：

$$h_{w,b}(x) = g(w^T x + b) \tag{12}$$

而判别函数为：

$$g(z) = \begin{cases} 1, & z \geq 0 \\ -1, & z < 0 \end{cases} \tag{13}$$

则求解最优线性分割器即为求解以下最优化问题：

$$\begin{aligned} \min_{w,b} \quad & \frac{1}{2} || w ||^2 \\ s.t. \quad & y^{(i)}(w^T x^{(i)} + b) \geq 1, i = 1, \ldots m \end{aligned} \tag{14}$$

由于存在噪音，一般情况下不可能线性完全可分，需要引入松弛变量，增加惩罚因子ζ_i，这里给出一阶软间隔分类器的形式，则优化形式为：

这里只是线性间隔吧，没有用到非线性的核方法

$$\begin{aligned} \min \quad & \frac{1}{2} || w ||^2 + C\sum_{i=1}^{I} \zeta_i \\ s.t. \quad & y_i[(wx_i) + b] \geq 1 - \zeta_i \\ & \zeta_i \geq 0 \end{aligned} \tag{15}$$

五、随机森林

随机森林方法是利用抽样，生成并建立多个随机的决策树，并将各个决策树汇总在一起进行分类判断的方法。关于决策树的建立方法这里不再详述，《The Elements of Statistical Learning》中对决策树有详细的介绍。

随机森林，其随机现在在构造单个决策树的两个方面。其在构造单个决策树时，首先随机地有放回地从 N 个建模样本中抽取 N 个样本出来，并随机地从 M 个特征中抽取 m（m<<M）个特征，由此进行决策树的构建。在重复地构建 t

个决策树后，决策森林即形成。在新的样本进入决策森林后，每一个子决策树都根据自己的规则对该样本进行判断，而后综合每一个决策树的结果综合判断。

第三章 实证模型

第一节 数据准备

本文使用的股吧文本数据由申万宏源研究所舆情系统提供。申万宏源研究所舆情系统的数据源主要分两个部分，一部分来自国内主流金融媒体的新闻报道，包括新浪财经、和讯财经、东方财富等网站的金融新闻；另一部分来自于国内最具影响力的股吧论坛的股评文本信息，即东方财富股吧的股评信息。东方财富是国内最大的金融媒体平台之一，而其东方财富股吧更是该公司的核心竞争产品，其股吧参与人数、信息传播速度等综合实力为行业第一，这里使用它的股吧文本信息作为研究对象较有研究意义和实践价值。

股吧文本数据库的具体构建方法为，按一只一只股票对东方财富股吧的每日股评进行网络爬虫爬取，记录的信息包括股票名、发言人昵称、发布时间、发布内容等相关信息。笔者为进行对股吧文本信息的情感判断研究，截取了 2015 年 12 月 1 日至 2015 年 12 月 7 日共 100 万条文本数据。为了进行情感标注，继续从这 100 万条数据中随机地抽取了 2000 条文本数据，并为这 2000 条信息进行情感标记，人工分为正向、负向两类情感。这样最后得到了 2000 条股吧评论信息，具体包括每条信息的内容和该条信息的情感方向。

在得到这 2000 条数据后，进一步地将数据随机切分成两组，其中 70%用于建模样本，30%用于验证样本。

第二节 文本特征选择

这里分词用的是之前提到的哪种方法？

在对建模样本进行分词分句之后，每一个观测，变为由一串单词组成的字符串。为了后续的数据挖掘工作，必须要将这些由字符组成的观测向量化。具体的做法就是，将特征单词进行验证，如果观测中出现了这个单词，就记为 1，如果没有出现这个单词就记为 0。下面就将通过特征选择的方法来找出有价值的特征信息。

如果选用NB的话，其实用词袋模型更好一点，即考虑一个词出现多次。不过这里如果先是进行特征选择的话，可以先用setofwords，仅记录是否出现

由于在文本中，会混含着‘我们’、‘你’、‘就’等对分类没有意义的所谓停用词，在这里首先对文本中所有的停用词进行剔除。作者所使用的停用词表也是知网提供的停用词词典，停用词词典的一部分如下：

表 3.1 知网停用词词典

也就是说	按照	比如	不独
啊	吧	鄙人	不管
阿	吧哒	彼	从
哎	把	彼此	从而
哎呀	罢了	边	打
哎哟	被	别	待
唉	本	不比	但
俺	本着	不成	但是
俺们	比	不单	当
按	比方	不但	当着

在进行停用词剔除之后，即可以进行特征词的选择。这里分别使用词频法、互信息法、卡方法三种方法对所有单词进行打分，并且选择得分最高的前 500 个单词选取为特征词。

没有使用信息增益法？

分别给出三种方法选出的前 20 个单词，如下：

表 3.2 特征词示例表

	词频法	互信息法	卡方方法
1	股	都	跑
2	今天	涨停	垃圾
3	都	股	出货
4	涨停	今天	散户
5	买	明天	买入
6	涨	垃圾	买进
7	明天	买	连续
8	主力	跌	涨停
9	跌	涨	快
10	垃圾	主力	傻
11	天	跑	太
12	大盘	散户	跌
13	大家	大家	银行
14	再	出货	大家
15	散户	大盘	明天
16	走	走	加油
17	大	买入	创新
18	跑	快	都
19	资金	天	加仓
20	元	太	试点

可见，不同方法筛选出来的特征单词在顺序上面有着非常显著的差异。在互信息法中出现的“都”排在第一位，而在卡方方法中却排在第 18 位；但是另一方面，虽然顺序上有着巨大的差别，但是选出来的单词却有一定的相似性。互信息法和卡方方法中有 50%以上的单词都是相同的。

不过关于不同特征选择方法的效果比较只能放在后面的情感判断结果中进行，观察到底哪一种特征选择方法更好。

第三节 情感判断建模和比较

在得到向量化表示的建模文本信息之后，分别使用情感词典方法、朴素贝叶斯方法、逻辑回归方法、支持向量机方法、随机森林方法进行建模，并将建模得到的参数估计结果使用到验证样本中对情感方向进行概率估计和方向判断。

为了对不同模型的估计效果进行比较，下面介绍三个对分类效果进行检验的指标。假设原始样本中有两个类别，其中正例个数为 P，负例个数为 N。而预测

分类中，有 TP 个样本被判为正例且实际为正例；FN 个样本被判为负例实际为正例；FP 个样本被判为正例实际为负例；TN 个样本被判为负例实际也是负例，如下表显示：

表 3.3 检验指标计算说明表

	判断为正例	判断为非正例
正例	TP	FN
非正例	FP	TN

由以上的列表，可以定义以下三个用于判断分类效果的指标：

（1）精确度(presicion)：P=TP/(TP+FP)，用于衡量在判断为正例的样本中到底有多少个是正例。

（2）召回率(recall)：R=TP/(TP+FN)，用于所有正例样本中，有多少是被判断为正例的。

（3）F1 measure：F=2*召回率*精确度/(召回率+精确度)，是召回率和精确度的一个综合指标。

一般来说，精确度和召回率在一定程度上是矛盾的。比如说，把待检测样本中全部判断为正例，那么该判断的召唤率就会达到 1，非常地高，但是此时的精确度就会降到很低的水平。而分类的目标当然是希望精确度和召回率都越高越好，所以才有了这么 F1 这种指标将两者综合起来。三个指标的范围都是 0 到 1 之间。以上三种指标被经常使用于对文本分类判别的效果比较，如 Yiming 和 Xin（1999）在文本分类方法比较时就使用了这种方法；而 Fujino，Isozaki 和 Suzuki（2008）使用 F1 score 方法来找最优的文本分类组合方法。本文将着重考虑 F1 score 的结果，并结合其他两个准则对分类效果进行比较。

下面给出五种方法在验证样本上的分类结果：

（1） 使用词频法特征选择方法

表 3.4 词频法分类效果检验

		presicion	recall	f1-score
情感词典	看多	0.62	0.63	0.63
	看空	0.57	0.57	0.57
	加权	0.65	0.6	0.6
朴素贝叶斯	看多	0.67	0.85	0.75
	看空	0.76	0.53	0.62
	加权	0.71	0.7	0.69

逻辑回归	看多	0.7	0.82	0.76
	看空	0.75	0.6	0.66
	加权	0.72	0.72	0.72
支持向量机	看多	0.71	0.78	0.74
	看空	0.72	0.63	0.67
	加权	0.71	0.71	0.71
随机森林	看多	0.65	0.75	0.69
	看空	0.65	0.54	0.59
	加权	0.65	0.65	0.64

（2）使用互信息法特征选择方法

表 3.5 互信息法分类效果检验

		presicion	recall	f1-score
情感词典	看多	0.61	0.64	0.62
	看空	0.58	0.55	0.56
	加权	0.59	0.6	0.59
朴素贝叶斯	看多	0.68	0.9	0.77
	看空	0.83	0.54	0.65
	加权	0.75	0.72	0.71
逻辑回归	看多	0.69	0.81	0.75
	看空	0.75	0.6	0.67
	加权	0.72	0.71	0.71
支持向量机	看多	0.72	0.79	0.75
	看空	0.74	0.66	0.69
	加权	0.73	0.73	0.73
随机森林	看多	0.68	0.7	0.69
	看空	0.66	0.63	0.65
	加权	0.67	0.67	0.67

（3）使用卡方方法特征选择方法

表 3.6 卡方方法分类效果检验

		presicion	recall	f1-score
情感词典	看多	0.63	0.68	0.65
	看空	0.61	0.54	0.57
	加权	0.62	0.62	0.62
朴素贝叶斯	看多	**0.68**	**0.89**	**0.77**
	看空	**0.81**	**0.54**	**0.65**
	加权	**0.74**	**0.72**	**0.71**
逻辑回归	看多	**0.72**	**0.83**	**0.77**
	看空	**0.78**	**0.64**	**0.7**
	加权	**0.75**	**0.74**	**0.74**
支持向量机	看多	**0.73**	**0.81**	**0.77**

	看空	**0.76**	**0.67**	**0.71**
	加权	**0.75**	**0.75**	**0.74**
随机森林	看多	0.72	0.66	0.69
	看空	0.65	0.71	0.68
	加权	0.69	0.68	0.69

首先，简单比较三种特征选择方法的分类效果。表中加灰部分的数字为加权的 recall，其实就是所有样本判断的正确率，即判断正确的样本/所有样本数，通过简单比较这个数字，就可以看到三种特征选择方法的优劣。可以发现，在卡方方法的下，五种情感判断的方法都比其他两种特征选择方法的正确率要高，可见卡方方法是明显优异于其他两种特征选择方法的。

鉴于卡方方法的效果最好，以下只看表 3 的结果。可以发现，相对于情感词典的方法，其他四种数据挖掘方法都明显要更好，而且好很多。

一般来说，从商业角度上来说，只有分类正确性达到 70%以上，就可以使用了。从这个水平来衡量，能达到这个标准的有朴素贝叶斯、逻辑回归、支持向量机三种方法，都可以满足这个要求。再进一步看具体的 presicion 值、recall 值以及 f1 值，可以发现，逻辑回归和支持向量机方法要明显比朴素贝叶斯方法效果好，而支持向量机方法和逻辑回归方法则非常接近。两者比较，支持向量机略微好一点。

我觉得如果用非线性的支持向量机结果可能更好

再继续进一步比较合格的三种方法在看多看空等细分方向的分类效果。不难发现，朴素贝叶斯方法相对其他两种方法有一个典型的特点，就是其在看多的方向上相对于看空的方向，presicion 明显要大而 recall 值明显要小，这种情况在其他两种方法上是不存在的。出现这种情况的原因，是因为朴素贝叶斯相对将过多的样本分类为看多造成的，而这又是由于朴素贝叶斯方法会使用到建模样本中多空的比例，并倾向于认为验证样本中依然会保持这个比例，而当验证样本中如果看多的样本相对较少时，就会出现这样的问题。而逻辑回归和支持向量机则没有这样的偏颇。可以看到，支持向量机和逻辑回归几乎在每一个点上的效果都是几乎一致的。

发现了朴素贝叶斯的一个特点

第四节 投资情绪与股票收益率关系研究初探

一、投资情绪指标构建

通过上述研究，获得了对股评信息进行情感判断的最优文本挖掘方法，接下来将通过以上的研究结论，对所有的股评信息进行情感判断，将每只股票的每个股评的情绪进行分类，并且进一步汇总并研究投资情绪和股票收益率的关系。

首先先给出接下来需要使用的情绪指标构建公式，单只股票的每月情绪指数为：

$$\text{月度负向情绪指标} = \text{负向情绪量} - \text{正向情绪量} \quad (16)$$

通过以上公式可以获得每只股票每个月的负向情绪指数。这里面之所以构建负向情绪指数，是因为股评信息中，负向情绪量一般都远大于正向情绪量。其中，负向情绪量，是给定月份的给定股票，所有负向股评的总数量；正向情绪量也是如此。

下面给出 2014 年 1 月份到 2015 年 8 月份的每月负向投资情绪的基本描述性统计：

表 3.7 月度负向投资情绪描述性统计

四分之一分位数	168710
中位数	257783
四分之三分位数	420229
均值	373695.8
标准差	462748.5

样本是指这20个月的情绪值。但是计算负面股评总数是不是不科学？因为这里评论总数可能是会变化的，不应该是选择负面评价比率更好吗

而月度负向投资情绪的概率密度函数图和累积分布图如下：

图 3.1　负向投资情绪概率密度分布图

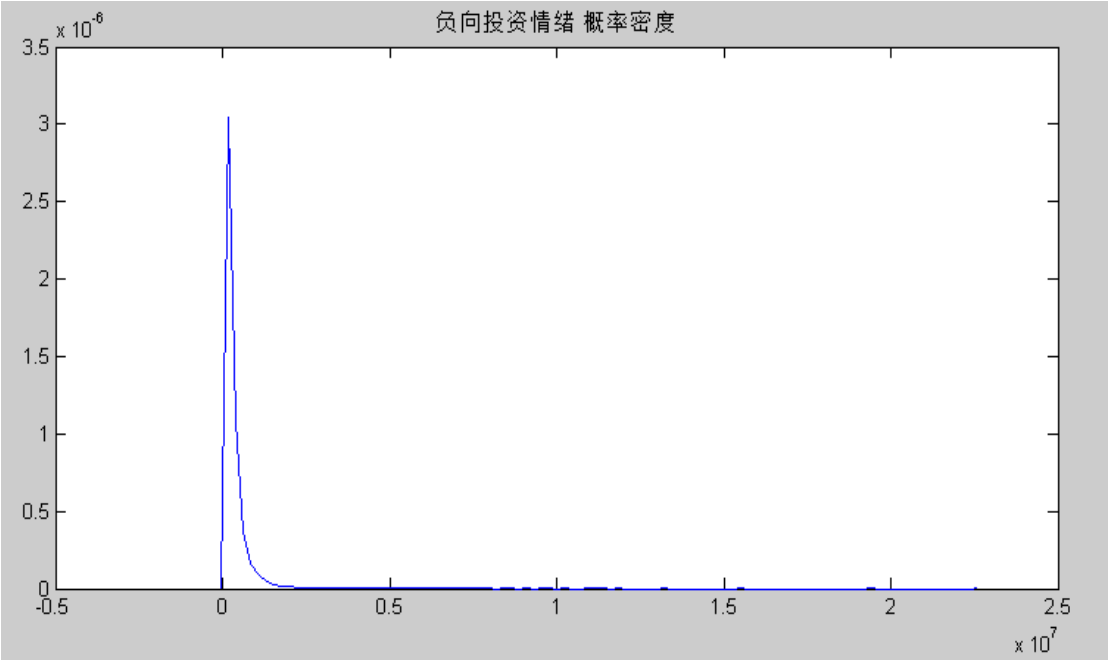
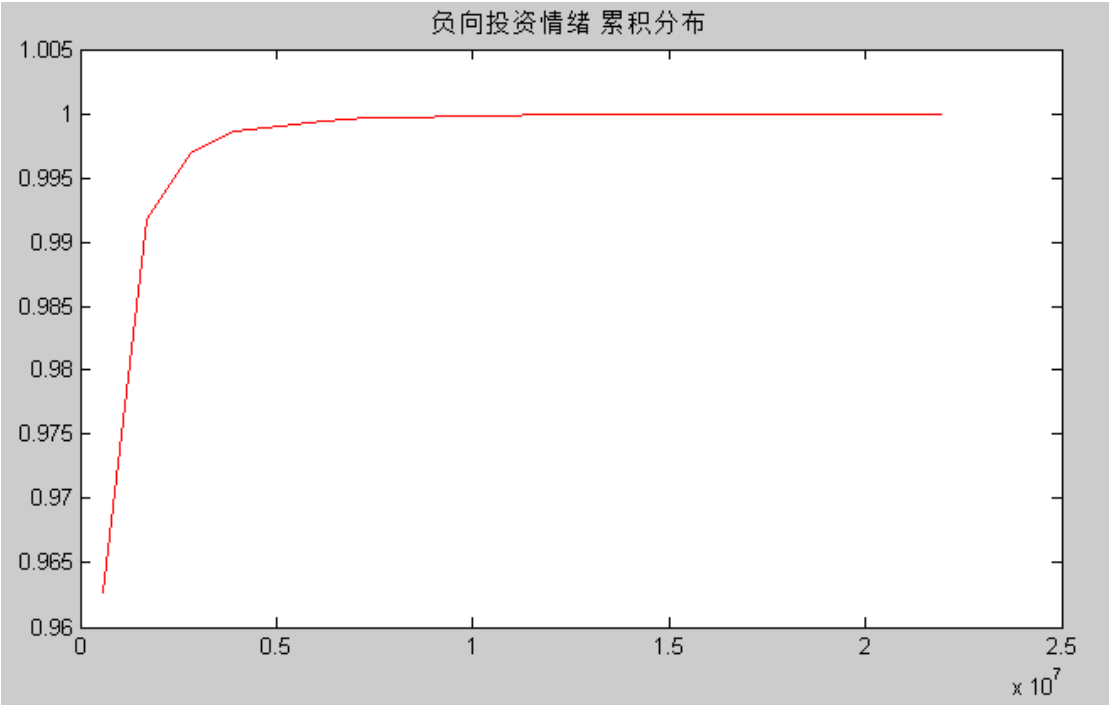


图 3.2 负向投资情绪累积分布图



由以上的统计性描述可以看到，投资情绪具有较大的偏度，大多数股票的月度负向情绪都集中在 0 到 500000 之间，而具有很少一部分的股票其负向投资情绪则能够达到很高的水平。

二、投资情绪和股票收益率关系研究

超额收益，alpha，还是不太明白，需要再去看一下

1、秩相关性分析

首先简单研究下投资情绪和股票收益率之间的相关关系。这里之所以进行的秩相关性分析，是由于相对于线性相关关系来说，秩相关关系在金融市场中的实践意义更大。在金融市场中，往往需要找出包含具体股票的投资组合，而这个投资组合能够包含正的超额收益即可，而不需要必须明确地规定收益率必须是多少。所以在股票池中，最实际的目标并不是预测收益率，而是要找出收益率相对高的股票。所以这里使用秩相关系数更符合实践经验。

注意，这里的投资者情绪和股票收益率分别是，本月的投资者情绪和下个月的股票收益率，即用当下可以看到的数据来预测未来的收益率，具备实际意义。

表 3.8　负向情绪指标和收益率相关性分析

月份	秩相关系数
2014年1月	-0.142446742
2014年2月	-0.164356237
2014年3月	-0.069373253
2014年4月	-0.031468653
2014年5月	-0.071192522
2014年6月	-0.008032535
2014年7月	-0.169281811
2014年8月	-0.185906885
2014年9月	-0.092813832
2014年10月	-0.038852126
2014年11月	0.246449144
2014年12月	-0.157339469
2015年1月	-0.043888866
2015年2月	-0.115745358
2015年3月	0.036370679
2015年4月	-0.191699119
2015年5月	-0.03691231
2015年6月	-0.049887187
2015年7月	-0.013508967

不应该是计算每天的更具实际意义吗？还是说因为构建投资组合持有时间比较长所以不要经常换。一个月内计算秩相关系数应该是按照不同股票做样本吧

选取了这么多月的情绪指数，但是之前说的只是截取了 2015 年12 月 1 日至 2015 年 12 月 7 日共 100 万条文本数据啊？答，前面的仅仅是训练模型，这里直接拿过来运用了。

从秩相关系数的结果中可以看到，在几乎全部的月份中，股票的下个月月度的收益率和本期的负向投资情绪都呈现负相关关系，除了 2014 年 11 月外，其他的秩相关系数都是负的。　在每一期的秩相关系数的基础上，再加入对这种关系的显著性检验，这里用的检验方法是 t 统计量，以下给出每个月的检验结果：

表 3.8　负向情绪指标和收益率相关性的显著性检验

月份	T检验P值
2014年1月	7.44E-13
2014年2月	1.15E-16
2014年3月	5.04E-04
2014年4月	1.15E-01
2014年5月	3.57E-04
2014年6月	6.87E-01
2014年7月	9.69E-18
2014年8月	3.24E-21
2014年9月	2.61E-06
2014年10月	4.91E-02
2014年11月	5.83E-37
2014年12月	8.45E-16
2015年1月	2.51E-02
2015年2月	2.80E-09
2015年3月	6.20E-02
2015年4月	1.16E-23
2015年5月	5.38E-02
2015年6月	8.69E-03
2015年7月	4.76E-01

如果针对一个股票而言，是不是就可以利用这个负向情绪指标来作为一个因子了？

从以上结果可以看出，除了 2014 年 4 月、2014 年 6 月和 2015 年 7 月，负向情绪指标和收益率之间都呈现了明显的相关性，可见两者之间的负向的关系是较为稳定而显著的。

后面这些金融的还没怎么看，感觉不太看得明白

2、排序打分方法研究投资情绪和股票收益率关系

在进行投资情绪和股票收益率的研究过程中，这里将采用金融市场中常用的排序打分方法，以下将详细介绍这种常用的排序打分构建投资组合的方法。

排序打分方法是金融行业中，研究股票某项指标和股票收益之前关系的重要方法，其具备的实践价值已经广为接受，无论是在美国、欧洲等高度发达的金融市场中，还是在国内正冉冉升起的量化投资领域，都是最重要的单因子收益能力检验的方法之一。下面将简述本文中，使用排序打分方法来考察负向投资情绪指标和股票收益率之间的关系的详细步骤：

- 回撤时间段为 2014.1-2015.8 月，调仓时点为每月的最后一个交易日
- 样本空间为在每个调仓时点去除掉上市不足 3 个月、ST、*ST 的所有 A 股
- 分组方法为每个调仓时点根据上个月投资情绪因子从小到大的顺序将样本

空间内这个月的非停牌股票分为 5 组，分别记为第 1 组（top 组合）……第 5 组（bottom 组合）

- 组合内运用行业内等权重加权计算每个组在每个月的组合收益率（即组合收益为组内所选个股收益的算术平均值），而行业间权重则使用该行业在中证 800 的权重。
- 构建投资情绪因子等权重多空组合；多空组合为 top-bottom 组合，即每月的等权重加权 top 组合-bottom 组合；
- 组合内运用行业内市值加权计算每个组在每个月的组合收益率（即组合收益为组内所选个股收益的算术平均值），而行业间权重则使用该行业在中证 800 的权重。
- 构建投资情绪因子多空组合；多空组合为 top-bottom 组合，即每月的市值加权 top 组合-bottom 组合；

通过以上的投资组合构建多空策略，并进行回测研究，研究实际投资效果，其回测参数为：交易成本约定为 2.5‰，即包括万分之五的佣金、千分之一的印花税以及千分之一的市场冲击；在获取回测结果之后，对每个投资组合进行评价，评价包括每个组合的月度净值收益计算年化收益率、年化超额收益、胜率、信息比、夏普比以及最大回撤。下面详细解释以上五个比较指标的含义：

年化收益率：即投资组合平均一年的投资收益率

年化超额收益：投资组合平均一年的收益率减去中证 800 一年的收益率

胜率：该投资组合收益率大于 0 的月数占总月数的比例

信息比：即年化收益率/收益率年化波动率

夏普比：（年化收益率-无风险利率）/收益率年化波动率

最大回撤：净值曲线从第一期开始，到最后一期，由最高点到最低点的损失比例。这个损失比例代表的意义非常重要，它代表的时候用这个投资组合进行投资可能出现的最坏情况，如果买在最高点（这是有可能的，且是最坏的情况），它要面临的损失程度。

首先给出中证 800 的收益情况，接下来的所有多空投资组合的比较基准都是这个投资组合。并且所有的投资组合，都是行业中性配置，即不同行业的股票的配权权重是中证 800 中每个行业的权重。

表 3.8 中证 800 基准收益情况

年化收益率	0.366707685
年化超额收益	0
夏普比	1.157358069
胜率	0
信息比	0
最大回撤	0.335213246

（1）行业间中性、行业内等权配置投资组合研究

接下来给出行业中性、行业内等权配置，各组股票的收益率表现结果，其中第一组是负向投资情绪最小的，故按其定义取名为 Bottom；而第五组为负向情绪最大的，为 Top 组合。

图 3.3 行业间中性行业内等权投资组合净值曲线

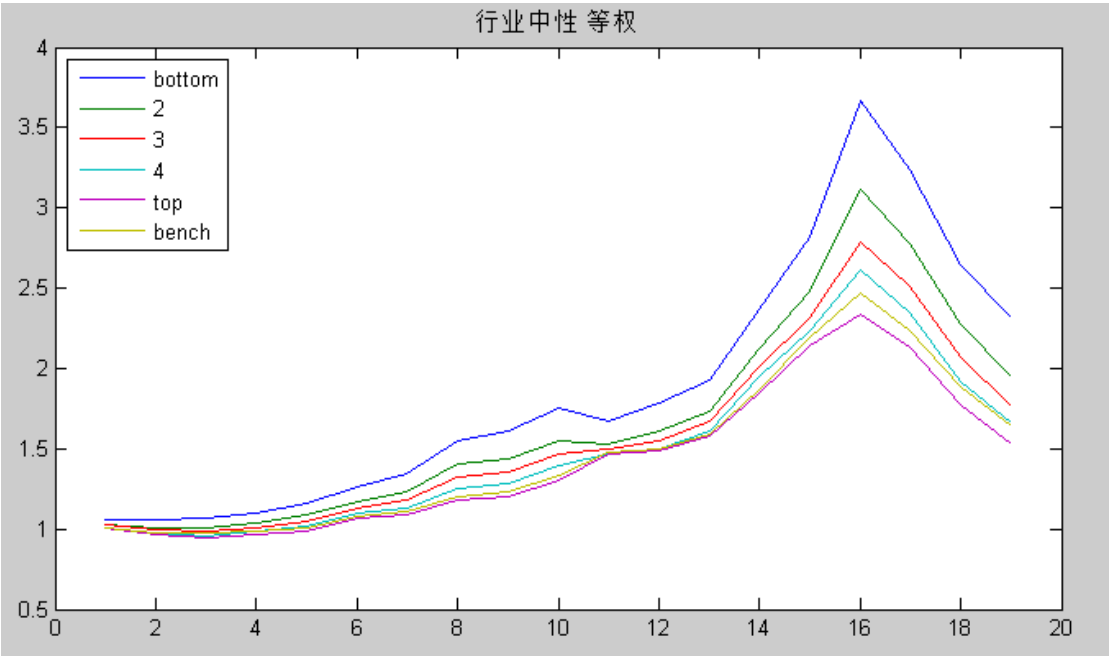


图 3.4 行业间中性行业内等权投资组合月度收益率汇总

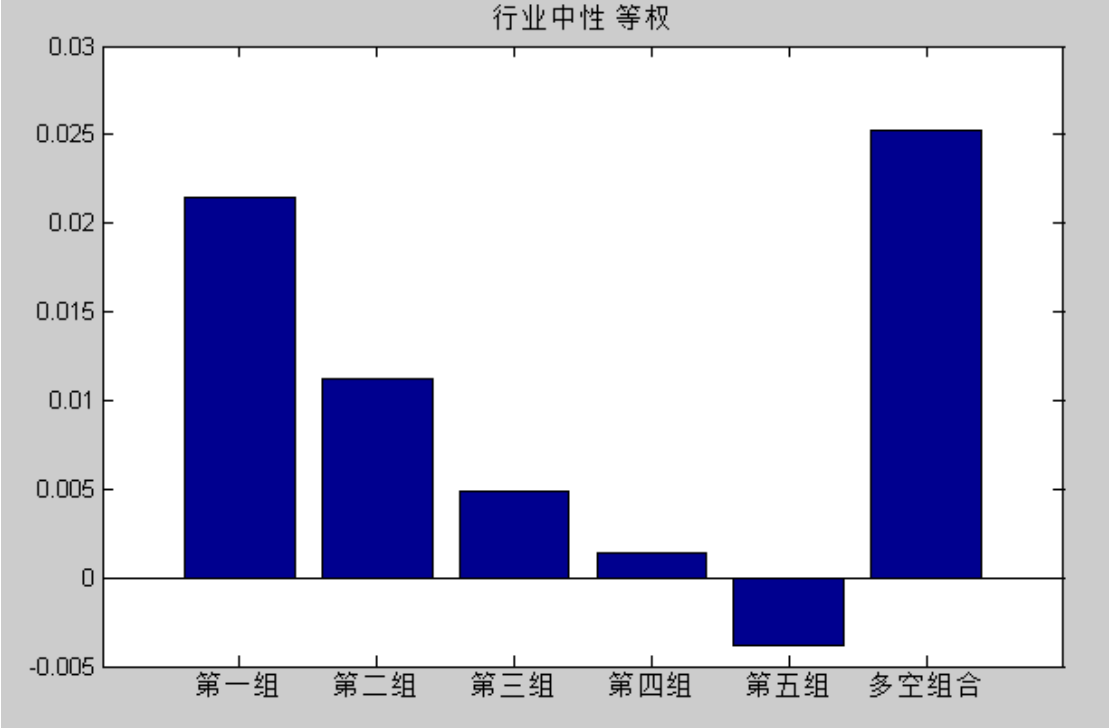


表 3.9　行业中性行业内等权投资组合收益评价表

	bottom	2	3	4	top
年化收益率	69.97%	52.25%	42.98%	37.73%	30.49%
年化超额收益	26.34%	12.89%	5.45%	1.42%	-4.56%
夏普比	8.27	9.06	9.85	10.21	11.07
胜率	0.84	0.68	0.63	0.63	0.37
信息比	1.23	0.82	0.53	0.21	-1.10
最大回撤	0.37	0.38	0.37	0.36	0.35

由以上图 3.4 和表 3.9 的情况可以看到，收益率在不同的负向投资情绪组别中呈现出明显的差异，并且这种差异是单调的，即负向投资情绪越小，则该投资组合的月度收益率越大。且多空组合（多第一组空第五组的投资组合）月度收益率可以达到 2.5%以上。

再结合图 3.3 可以看到，不同负向情绪的股票组合，其投资组合的净值曲线也是保持着稳定的差异。蓝色线，即 bottom 组合，也就是负向情绪最小的那个股票组合，其收益率净值曲线稳定地跑赢其他几个组合。

（2）行业间中性、行业内流通市值加权投资组合研究

再给出行业中性、行业内流通市值加权配置，各组股票的收益率表现结果：

图 3.5 行业间中性行业市值加权投资组合净值曲线

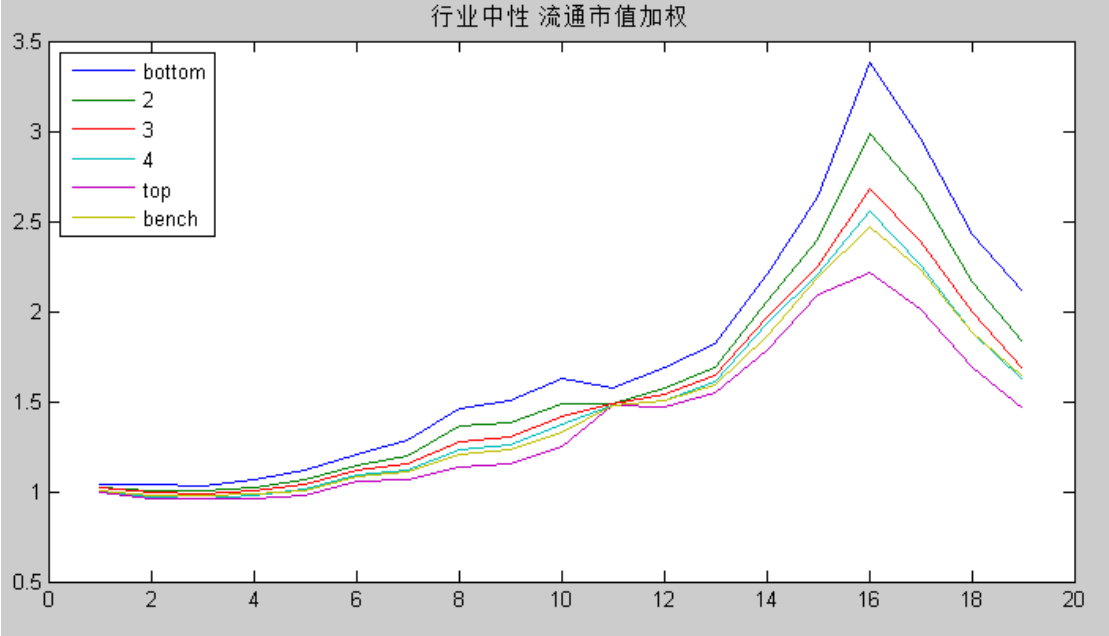


图 3.6 行业间中性行业市值加权投资组合月度收益率汇总

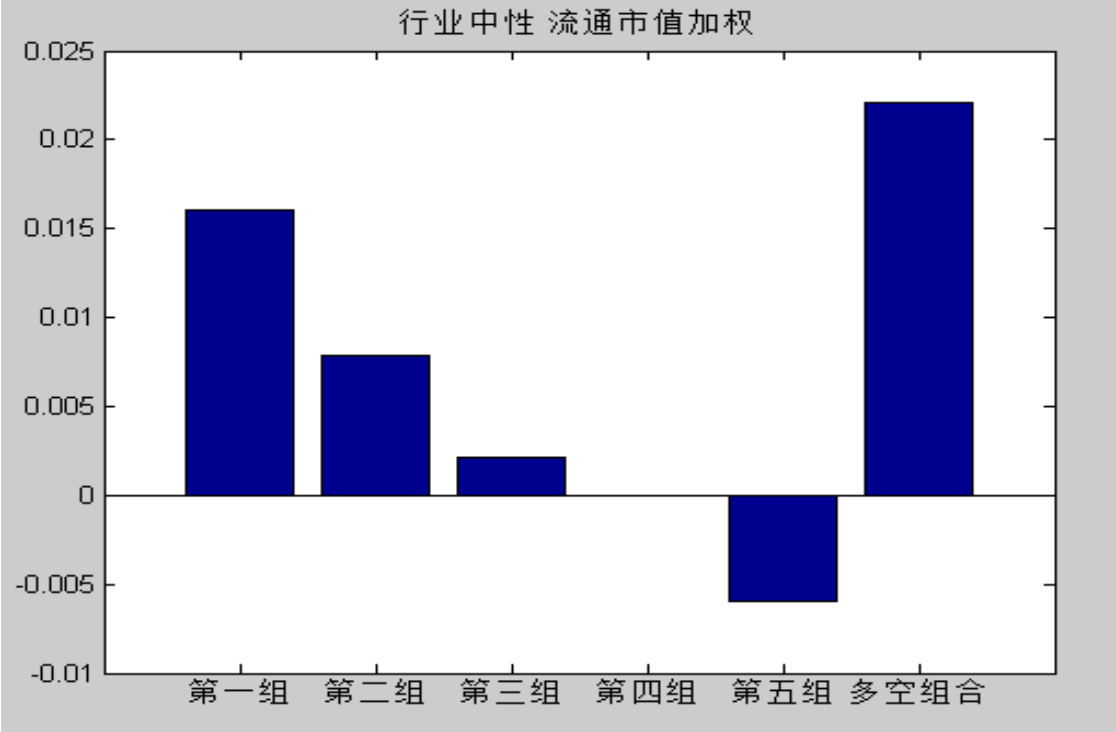


表 3.9 行业中性行业内流通市值加权投资组合收益评价表

	bottom	2	3	4	top
年化收益率	60.30%	46.56%	38.91%	35.74%	27.21%
年化超额收益	19.03%	8.67%	2.23%	-0.20%	-7.31%
夏普比	8.65	9.31	10.20	10.40	11.63
胜率	0.74	0.63	0.63	0.58	0.21
信息比	1.02	0.62	0.30	-0.01	-0.83
最大回撤	0.38	0.39	0.37	0.36	0.34

由于行业内等权配置方法，相当于在大市值股票和小市值股票配以相同的权重，其实本质上相当于。而在股票市场中，市值因素对股票收益率的影响较为显著，所以这里再次构建的投资组合是行业间中性构建投资，而行业内的股票用流通市值加权，由此得到五个投资组合。

由五组的结果可以看到，在行业中性行业内等权配置投资组合的两个结论这里依然成立，不过其月度平均收益率下降到0.23%左右的水平。这个比之前的略有下降，但是不同组之间收益的单调性和明显的差异依然存在。

第四章、研究结论及展望

第一节 研究结论

在这篇文章中，作者使用东方财富股吧的股评信息进行情感判断的研究，试图找到最适合中国金融市场股票评论的情感判断方法。

通过尝试三种特征选择方法，即词频法、互信息法、卡方方法后发现，卡方方法要明显优于另外两种特征选择方法。进而比较五种不同的情感分类器，包括情感词典方法、朴素贝叶斯方法、逻辑回归方法、支持向量机、随机森林。其中朴素贝叶斯方法、逻辑回归方法、支持向量机三种方法能够达到70%以上的准确性，可以投入使用；而其中支持向量机的分类效果相对最好，准确度能够达到75%，是最有效的方法。因此，笔者认为，在中国金融市场上，使用卡方方法进行特征选择，并且使用支持向量机进行情感分类的方法是相对来说较有效的文本挖掘方法。

第二节 研究展望

在文章的最后，初步探讨了投资情绪和股票收益率之间的关系，可以看到投资情绪对未来投资收益有一定程度的影响。在以后的研究中，非常值得进一步地探讨股票未来收益率和投资情绪的关系，甚至包括情绪变化、情绪波动的等等的关系；还可以从波动率的角度，研究股票波动率和投资情绪的潜在联系。最后，从商业实践的角度，还可以将投资者情绪加入多因子选股模型，对构建投资组合、建立量化投资策略贡献帮助。

参考文献

[1]Wysocki P D. Cheap talk on the web: The determinants of postings on stock message

[2]boards[J]. University of Michigan Business School Working Paper, 1998 (98025).

[3]Antweiler W, Frank M Z. Is all that talk just noise? The information content of internet stock message boards[J]. The Journal of Finance, 2004, 59(3): 1259-1294.

[4]Da Z, Engelberg J, Gao P. In search of attention[J]. The Journal of Finance, 2011, 66(5): 1461-1499.

[5]Joseph K, Wintoki M B, Zhang Z. Forecasting abnormal stock returns and trading volume using investor sentiment: Evidence from online search[J]. International Journal of Forecasting, 2011, 27(4): 1116-1127.

[6]Antweiler W, Frank M Z. Is all that talk just noise? The information content of internet stock message boards[J]. The Journal of Finance, 2004, 59(3): 1259-1294.

[7]Tetlock P C. Giving content to investor sentiment: The role of media in the stock market[J]. The Journal of Finance, 2007, 62(3): 1139-1168.

[8]Bozic C, Seese D. Neural networks for sentiment detection in financial text[C]//Proceedings of the 14th International Business Research Conference. 2011.

[9]Bollen J, Mao H, Zeng X. Twitter mood predicts the stock market[J]. Journal of Computational Science, 2011, 2(1): 1-8.

[10]Ormos M, Vázsonyi M. Impacts of Public News on Stock Market Prices: Evidence from S&P5001[J]. Interdisciplinary Journal of Research in Business, 2011, 1(2): 01-17.

[11]赵丽丽，赵茜倩，杨娟，等．财经新闻对中国股市影响的定量分析[J]．山东大学学报（理学版），2012，47(7)：70-75.

[12]祝宇．网络信息对于股票市场的影响[D]．浙江大学，2013.

[13]范辛亭,庄皓亮,杨靖凤 ．事件选股策略之新闻选股．长江证券研究报告．2014-2-14

[14]Friedman J, Hastie T, Tibshirani R. The elements of statistical learning[M]. Springer, Berlin: Springer series in statistics, 2001.

[15]Yang Y, Liu X. A re-examination of text categorization methods[C]//Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 1999: 42-49.

[16]Fujino A, Isozaki H, Suzuki J. Multi-label Text Categorization with Model Combination based on F1-score Maximization[C]//IJCNLP. 2008: 823-828.

[17]Wysocki P D. Cheap talk on the web: The determinants of postings on stock message boards[J]. University of Michigan Business School Working Paper, 1998 (98025).

[18]Schumaker R P, Chen H. Textual analysis of stock market prediction using breaking financial news: The AZFin text system[J]. ACM Transactions on Information Systems (TOIS), 2009, 27(2): 12.

[19] Choi H, Varian H. Predicting the present with Google Trends[J]. Economic Record, 2012, 88(s1): 2-9.

[20] Kim O, Verrecchia R E. Market reaction to anticipated announcements[J]. Journal of Financial Economics, 1991, 30(2): 273-309.

[21] Wilson T, Wiebe J, Hoffmann P. Recognizing contextual polarity in phrase-level sentiment analysis[C]//Proceedings of the conference on human language technology and empirical methods in natural language processing. Association for Computational Linguistics, 2005: 347-354.

[22] Asur S, Huberman B A. Predicting the future with social media[C]//Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on. IEEE, 2010, 1: 492-499.

[23] O'Connor B, Balasubramanyan R, Routledge B R, et al. From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series[J]. ICWSM, 2010, 11(122-129): 1.2.

[24] Baker M, Wurgler J. Investor sentiment and the cross‐section of stock returns[J]. The Journal of Finance, 2006, 61(4): 1645-1680.

[25] Barber B M, Odean T. All that glitters: The effect of attention and news on the buying behavior of individual and institutional investors[J]. Review of Financial Studies, 2008, 21(2): 785-818.

[26] Boehmer E, Jennings R, Wei L. Public disclosure and private decisions: Equity market execution quality and order routing[J]. Review of Financial Studies, 2007, 20(2): 315-358.

[27] Hirshleifer D, Teoh S H. Limited attention, information disclosure, and financial reporting[J]. Journal of accounting and economics, 2003, 36(1): 337-386.

[28] Brown S J, Warner J B. Using daily stock returns: The case of event studies[J]. Journal of

financial economics, 1985, 14(1): 3-31.

[29] Rajgopal S, Kotha S, Venkatachalam M. The relevance of web traffic for internet stock prices[M]. Graduate School of Business, Stanford University, 2000.

[30] Yang Y. An evaluation of statistical approaches to text categorization[J]. Information retrieval, 1999, 1(1-2): 69-90.

[31] Lee C, Shleifer A, Thaler R H. Investor sentiment and the closed‐end fund puzzle[J]. The Journal of Finance, 1991, 46(1): 75-109.

致谢

光阴似箭，时光荏苒，转眼间忙碌的研究生生涯就走到了尽头。回想在母校生活的六年中，我的心里充满了感慨，而对母校的感恩之情也溢于言表。感谢母校对我的培养，为我们提供这么优良的学习资源、生活环境；感谢母校对我的谆谆教诲，“厚德博学，经济匡时”的校训将伴我一生；感谢两年以来陪我走过风风雨雨的研究生同学、老师、朋友，你们的鼓励和帮助都让我受益匪浅。

首先衷心感谢导师骆司融老师，正是在他的悉心的指导下本文才能够最终顺利完成。他严谨的治学态度、精益求精的学术精神、认真勤奋的工作态度，都让我十分敬佩，并且也是我以后的生活和工作中的榜样。

感谢在申万宏源实习期间的同事，金融工程分析师王小川老师，其在我的研究工作中提供了巨大的帮助，为我提供非常难得的股吧信息数据源，并给予我方法论上的指导和帮助，是我完成论文不可缺少的重要基础。

感谢 2014 应用统计硕士班的所有同学，我们一起成长，一起进步，将来将一起踏入工作岗位，一起追求梦想。在财大有你们陪伴的每一天，我都会牢牢记在心里。

感谢家人的养育之恩，愿你们幸福快乐每一天。

再次感谢每一位帮助过我的老师、同学、朋友，对你们致以最诚挚的谢意。

孔潇

二〇一六年三月于上海财经大学

附录

附录 1 不同特征选择选出的特征词

词频法			互信息法			卡方方法		
股	见	基金	都	以上	开启	是不是	跑	天天
今天	抢	为证	涨停	发展	持续	举报	垃圾	明显
都	一八三	忽悠	股	尾盘	业绩	拉高	出货	呵呵
涨停	增发	红	今天	准备	制造	狗庄	散户	反弹
买	收购	复牌	明天	前	鼓掌	断崖	买入	不能
涨	以后	疯狂	垃圾	暴涨	价	SB	买进	下跌
明天	万股	掉	买	上涨	一天	中信证券	连续	不涨
主力	证券	留下	跌	大单	买点	操	涨停	机构
跌	到底	合作	涨	分	为证	绿	快	以上
垃圾	一起	骗	主力	一点	复牌	逃	傻	发展
天	流出	喊	跑	一定	合作	换入	太	尾盘
大盘	持有	倍	散户	钱	倍	向下	跌	下午
大家	妖	不到	大家	到底	不到	跳	大家	被套
再	洗盘	期	出货	流出	期	新城	明天	滴
散户	特力	几个	大盘	一下	牛股	半年	加油	难以
走	交易	千万别	走	长期	力	圾	创新	臭
大	四	牛股	买入	投资	住	不敢	都	懂
跑	加仓	亏损	快	收复	趋势	大量	加仓	全都
资金	打压	后悔	天	有人	最后	割	试点	负
元	试点	好股	太	兄弟	开盘	鉴定	亿元	幸亏
没有	亿元	全部	跌停	SDR	最好	屎	解冻	多点
出货	进入	向上	没	政策	吃	两名	外资	同样
没	接盘	打开	再	成功	打新	历史	八日	除权
好	摊手	一批	大	未来	原因	每天	连阳	一涨

说	解冻	次新股	资金	红利	板	所有	垃圾股	美的
跌停	融资	上升	元	涨停板	第一	却	托	st
买入	外资	长	说	回	号	注意	逼	周一
卖	八日	猪	好	重大	追	海	开展	长江
股票	连阳	追高	去	两	升	看好	狂飙	明白
月	第一	直接	盘	期待	小	会	收涨	东方
快	开展	周三	买进	建仓	越来越	一步	掀起	开创
看	一步	股灾	没有	赚钱	个股	信心	低价	自摸
太	号	国际	利好	短线	时间	坚持	地产股	命
利好	量	新高	傻	左右	流入	转	护盘	瞅
会	信心	实力	连续	上去	说明	节奏	小幅	庄托
想	追	破股	人	希望	收	市场	三	狗日
人	变	电视	逼	还要	分析	下周	指大涨	工程
不要	坚持	卖掉	卖	k	洗	增持	助沪指	撤
重组	升	明日	看	倒	早盘	到位	满仓	阴线
下午	力	发行	不要	大便	货	吸	去	坑里
已经	住	即将	已经	没人	早	底部	盘	烂
机会	趋势	找	拉	兴奋	压盘	尼	大笑	多月
去	近期	反正	想	吃屎	看来	三大	走	人气
盘	最后	挂	机会	千万	集团	美联储	抛	老鼠
拉	B	吸引	会	全	跳水	计划	跌停	股权
应该	有点	下降	月	解套	朋友	上午	不赞	恶意
死	股市	下面	股票	公告	股东	放缓	小心	投
股价	昨天	恶庄	下午	高点	块	入篮	改革	原
公司	小	日线	应该	式	信	稳定	大盘	获利
上	开盘	进来	做	一步	基金	勇攀	妈	看到
砸	最好	就要	玩	信心	忽悠	激动	真是	能源
很	越来越	横盘	出来	坚持	红	搬家	接盘	跌下来
拉升	转	磨叽	出	转	疯狂	肉	摊手	弱势

股份	肯定	一个月	重组	节奏	掉	收钱	利好	格力电器
逼	亿	支撑	死	市场	留下	爆发	长期	不争气
不是	节奏	涨幅	砸	下周	骗	估值	投资	伤心
出来	吃	华发	抛	增持	喊	六	收复	盘子
买进	狂飙	基本	拉升	到位	几个	出现	有人	气
看看	个股	走强	几天	以下	千万别	新药	兄弟	家里
只	打新	内	大笑	低开	吸	体育	SDR	想想
出	怒	搞	哭	马上	底部	概率	政策	指数
现在	收涨	大胆	上	时	尼	实属	成功	早就
不会	时间	炒股	看看	最	三大	推荐	未来	必跌
筹码	不行	券商	现在	新股	美联储	胜利	红利	卖光
继续	掀起	表演	创业板	科技	计划	农业	涨停板	做空
做	跟着	董事长	A股	一八三	上午	正	回	放心
傻	缺爹	久	妈	收购	放缓	坚守	重大	手贱
个点	原因	低吸	真是	万股	入篮	坚定	两	反对票
哭	市场	国企	股价	证券	稳定	大户	期待	电话
玩	狗	重新	公司	一起	勇攀	底	建仓	操纵
后	低价	操盘手	线	妖	激动	加	做	抛售
连续	地产股	确定	卖出	特力	搬家	加注	玩	经验
进	仓	哥	很	起来	肉	牛市	龙头	换
赚	别人	不想	不是	量	收钱	处于	净流入	五个
高	护盘	强势	垃圾股	变	爆发	预测	成本	玩吧
日	难道	成为	托	有点	估值	第一股	没	开玩笑
庄	小幅	位置	庄	股市	六	影视	拉	耍
调整	下周	老子	不会	昨天	出现	捡便宜	创业板	怎么回事

创业板	三	里面	继续	肯定	新药	势	A 股	过来
几天	票	东	股份	不行	体育	这次	近期	上演
A 股	增持	深	只	跟着	概率	背书	B	TM
行情	流入	均线	加油	狗	实属	概念股	亿	心里
下	指大涨	客	创新	仓	推荐	维稳	怒	一半
线	说明	顶	抄底	票	胜利	股友	缺爹	出局
卖出	收	重工	低	亏	农业	补涨	别人	四个
大笑	到位	盘面	大跌	放	正	跟进	难道	选
庄家	助沪指	炒	小散	走势	坚守	有意	几天	当天
抛	满仓	空间	改革	一次	坚定	集体	人	关系
大涨	亏	解禁	筹码	此股	大户	耐心	板块	中信
抄底	放	暂时	个点	震荡	底	主线	动能	配合
可能	分析	嘿嘿	不赞	问题	加	召开	智能	折磨
割肉	板	决定	小心	这种	加注	挂单	部署	贴
低	走势	下来	后	这股	牛市	下周一	味精	合并
妈	还要	终于	高	强	处于	目标	飞	海油
看好	k	短期	板块	见	预测	后市	稳住	黑
知道	赶紧	挣	进	以后	第一股	尖	牛	吃饱
真是	谢谢	真	赚	交易	影视	感觉	长线	金
家	长期	立帖	看好	动能	捡便宜	集合竞价	不错	清仓
地产	一波	成	大涨	智能	势	不该	坚决	一天到晚
大跌	倒	完毕	加仓	部署	这次	监管	上攻	减仓
小散	投资	封	试点	味精	背书	赞	仓位	清
新	买回来	封板	亿元	飞	概念股	必有	实质性	去死吧
天天	大便	层	解冻	稳住	维稳	提振	介入	这货
明显	周五	动能	外资	牛	股友	马上拉	踩	滴汗
收盘	没人	智能	八日	长线	补涨	数据	上车	好于

呵呵	收复	带走	连阳	不错	跟进	发动	暴风	道
反弹	发财	上市	套	坚决	有意	力荐	行业	单
不能	进去	完	暴跌	上攻	集体	献花	老夫	说声
下跌	比较	部署	日	仓位	耐心	陪	放量	死庄
准备	兴奋	逃命	调整	实质性	主线	布局	全球	坑
前	这么久	拜神	行情	介入	召开	虹普	莲花	装
不涨	有人	味精	下	踩	挂单	绝对	电器	奉劝
机构	兄弟	管理人员	天天	上车	下周一	电气	完成	关注
板块	预期	跌幅	明显	暴风	目标	接回来	不信	错
回来	吃屎	飞	呵呵	行业	后市	守住	国务院	严重
停牌	SDR	直	反弹	老夫	尖	预计	珠峰	彻底
改革	动	稳住	不能	放量	感觉	资产	潮	无语
垃圾股	千万	牛	下跌	全球	集合竞价	选择	抢	确实
左右	送	倒闭	不涨	莲花	不该	军工	增发	何必
上去	政策	长线	机构	电器	监管	抢筹	持有	运气
希望	能否	不错	割肉	完成	赞	地区	洗盘	总
套	成功	老	龙头	不信	必有	仙人指路	四	劝
暴跌	未来	投资者	净流入	国务院	提振	功能	套	请
托	红利	是不是	成本	珠峰	马上拉	厚报	暴跌	硬
今日	价格	举报	接盘	打压	数据	翻倍	抄底	热闹
暴涨	回调	拉高	摊手	进入	发动	全仓	低	好好
一下	涨停板	坚决	可能	融资	力荐	高开	还要	以前
上涨	良心	狗庄	两天	带走	献花	三联	k	茅坑
估计	全	断崖	庄家	上市	陪	绝不	倒	利用
两天	解套	SB	家	完	布局	四十	大便	开
才	方案	中信证券	地产	逃命	虹普	拉到	没人	四九
龙头	公告	操	新	拜神	绝对	分红	兴奋	走坏

加油	回	绿	唱	管理人员	电气	使劲	吃屎	潍柴
净流入	少	逃	不好	跌幅	接回来	分类	千万	不再
不赞	重大	上攻	已	直	守住	条件成熟	全	缺德
唱	一路	换入	开展	倒闭	预计	医药	解套	贱
赚钱	高点	仓位	狂飙	老	资产	意思	公告	不动
以下	很多	实质性	收涨	投资者	选择	附股	高点	两市
以上	影响	向下	掀起	是不是	军工	家化	式	紧
低开	两	跳	低价	举报	抢筹	潜力	机会	一跌
创新	式	新城	地产股	拉高	地区	石墨	线	墨迹
发展	行	半年	护盘	狗庄	仙人指路	长远	卖出	转前
马上	坐等	圾	小幅	断崖	功能	获	出来	相当
时	期待	不敢	三	SB	厚报	弱股	出	尼玛
小心	快点	介入	指大涨	中信证券	翻倍	启动	好	纠结
最	建仓	大量	助沪指	操	全仓	很快	两天	觉得
短线	盈利	踩	满仓	绿	高开	吸货	元	疼
不好	洗	上车	知道	逃	三联	沈机	拉升	阴跌
起来	早盘	暴风	今日	换入	绝不	电力	唱	受
成本	开启	行业	估计	向下	四十	领域	不好	该死
已	持续	割	才	跳	拉到	翻番	已	道理
尾盘	业绩	老夫	回来	新城	分红	超级	哭	帖子
一次	制造	鉴定	停牌	半年	使劲	四季度	大跌	上证
潮	鼓掌	放量	潮	圾	分类	值得	小散	晕
大单	货	全球	抢	不敢	条件成熟	便宜	应该	问候
新股	早	莲花	增发	大量	医药	储蓄	说	舒服
此股	压盘	屎	持有	割	意思	不要	带走	小盘股

分	价	两名	洗盘	鉴定	附股	已经	上市	看看
一点	看来	历史	四	屎	家化	资金	完	现在
震荡	集团	每天	近期	两名	潜力	庄	逃命	今天
一定	跳水	电器	B	历史	石墨	大单	拜神	大涨
问题	朋友	所有	亿	每天	长远	分	管理人员	想
科技	一天	完成	怒	所有	获	一点	跌幅	盈利
钱	股东	不信	缺爹	却	弱股	一定	直	开启
这种	块	国务院	别人	注意	启动	钱	倒闭	持续
这股	信	却	难道	海	很快	到底	老	业绩
强	买点		收盘	盈利		流出	投资者	

附录 2 核心代码——分词分句（python）

```
# 分句分词
guba_comment_split_words = []
quit_loc = []
tick = 0
for comment_long_sent in guba_comment:
    try:
        comment_short_sent = cut_sentence_to_sent(comment_long_sent)
    except:
        comment_split_words = []
        guba_comment_split_words.append(comment_split_words)
        quit_loc.append(tick)
        tick = tick + 1
        continue
```

```
        comment_split_words = []
        for short_sentence in comment_short_sent:
            comment_split_words = comment_split_words +
cut_sent_to_word(short_sentence)

        guba_comment_split_words.append(comment_split_words)
        tick = tick + 1

# 把短句分成单词
def cut_sent_to_word(words):
        seg = jieba.cut(words)
        split_words = []
        for i in seg:
            split_words.append(i)
        return split_words

# 把长句分成短句
def cut_sentence_to_sent(words):
        # words = (words).decode('utf8')
        start = 0
        i = 0
        sents = []
        punt_list = ',.!?:;~，。！？：；～[]'.decode('utf8')

        # 先保证第一个字符不是标点符号
        start_loc = 0
        for word in words:
            if word in punt_list:
                start_loc = start_loc + 1
```

```
            else:
                continue
        words_adjust = words[start_loc:]


        for word in words_adjust:
            if word in punt_list and token not in punt_list: #检查标点符号下一个字符是否还是
标点
                sents.append(words_adjust[start:i])
                start = i+1
                i += 1
            else:
                i += 1
                token = list(words_adjust[start:i+2]).pop() #  取下一个字符
        if start < len(words_adjust):
            sents.append(words_adjust[start:])
        return sents
```

附录 3 核心代码——特征选择（python）

```
# 通过特征选择方法选择特征词
pos_words = []
neg_words = []


for i in range(len(guba_attitude_train_adjust)):
    if guba_attitude_train_adjust[i] == 0:
        continue
    if guba_attitude_train_adjust[i] == 1:
        pos_words = pos_words + guba_comment_train_adjust[i]
    else:
        neg_words = neg_words + guba_comment_train_adjust[i]
```

```
# 计算信息量
word_fd = FreqDist(pos_words+neg_words + posBigrams + negBigrams)
cond_word_fd = ConditionalFreqDist()
cond_word_fd['pos'].update(pos_words + posBigrams)
cond_word_fd['neg'].update(neg_words + negBigrams)

pos_word_count = cond_word_fd['pos'].N() #积极词的数量
neg_word_count = cond_word_fd['neg'].N() #消极词的数量
total_word_count = pos_word_count + neg_word_count

word_scores = {}
for word, freq in word_fd.iteritems():
    pos_score = BigramAssocMeasures.chi_sq(cond_word_fd['pos'][word], (freq,
pos_word_count), total_word_count) #计算积极词的卡方统计量
    neg_score = BigramAssocMeasures.chi_sq(cond_word_fd['neg'][word], (freq,
neg_word_count), total_word_count) #同理
#     pos_score = BigramAssocMeasures.raw_freq(cond_word_fd['pos'][word], (freq,
pos_word_count), total_word_count) #计算积极词的卡方统计量
#     neg_score = BigramAssocMeasures.raw_freq(cond_word_fd['neg'][word], (freq,
neg_word_count), total_word_count) #同理
#     pos_score = BigramAssocMeasures.mi_like(cond_word_fd['pos'][word], (freq,
pos_word_count), total_word_count) #计算积极词的卡方统计量
#     neg_score = BigramAssocMeasures.mi_like(cond_word_fd['neg'][word], (freq,
neg_word_count), total_word_count) #同理


    word_scores[word] = pos_score + neg_score #一个词的信息量等于积极卡方统计量
加上消极卡方统计量
```

```
# 去除停用词之后保留信息量最大的 N 个单词作为特征信息
number = 500
stop_words = get_txt_data('data_cache/stop_words.txt', 'lines') + \
                get_txt_data('data_cache/stop_words_guba.txt', 'lines')
feature_words = find_best_words(word_scores, stop_words ,number )

def find_best_words(word_scores, stop_words ,number ):
    sorted_vals = sorted(word_scores.iteritems(), key=lambda (w, s): s, reverse=True) #把词按信息量倒序排序。number 是特征的维度，是可以不断调整直至最优的
    tick = 0
    best_words = []
    for w,s in sorted_vals:
        if w not in stop_words:
            best_words.append(w)
            tick = tick + 1
        if tick == number:
            break
    return best_words
```

附录 4 核心代码——情感判断（python）

```
def class_model_for_guba(classifier,train,testSet,tag_test):
    classifier = SklearnClassifier(classifier) #在 nltk 中使用 scikit-learn 的接口
    classifier.train(train) #训练分类器

    pred = classifier.classify_many(testSet) #对开发测试集的数据进行分类，给出预测的标签
    target_names = ['pos','neg']
```

```
    return classification_report(tag_test, pred,target_names=target_names)
```