# 基于微博的股票投资者未来情感倾向识别研究

庞 磊 李寿山 张 慧 周国栋

（苏州大学计算机科学与技术学院 苏州 215006）（苏州大学自然语言处理实验室 苏州 215006）

**摘 要** 近年来，微博越来越受到网络用户的青睐，成千上万的用户通过发布微博共享他们的观点和情感。其中，有大量带有情感倾向（认为某事物“好”或“坏”）的微博，这些微博反映了作者的情绪。投资者情绪（investor sentiment）是研究经济市场走向的重要指标，行为金融学认为股票投资者情绪影响投资者决策，进而影响股票市场，而反映股票投资者情绪的重要指标是投资者对股票市场未来行情的情感倾向（认为股票市场未来行情“好”或“坏”）。通过对新浪微博（目前最大的中文微博平台）上股票投资者发布的文本进行情感信息方面的分析与研究，提出了一种自动识别股票投资者未来情感倾向的方法。该方法分为两级识别，第一级是：识别出微博中包含未来情感的句子；第二级是：将第一级识别出来的包含未来情感的句子分为正面评论（看涨）和负面评论（看跌）。实验结果表明，所提方法对自动识别股票投资者的未来情感倾向达到了非常好的效果。

**关键词** 计算机应用，中文信息处理，投资者情绪，微博，情感分类，情感倾向

**中图法分类号** TP391 **文献标识码** A

## Method to Identify the Future Stock Investor Sentiment Orientation on Chinese Micro-blog

PANG Lei LI Shou-shan ZHANG Hui ZHOU Guo-dong

(School of Computer Science and Technology, Soochow University, Suzhou 215006, China)

(Natural Language Processing Lab, Soochow University, Suzhou 215006, China)

**Abstract** Recently, Micro-blog has attracted more and more interests of internet users. Thousands of the users share their views and opinions through micro-blog. There are a large number of texts with sentiment orientation (thinking something is “good” or “bad”) on the Micro-blog. These texts reflect the authors’ emotion. Investor sentiment is the important indicator to research on the economic market trends. On behavioral finance, Stock investor sentiment affects the investors’ decision. Then it will affect stock market. The stock investor sentiment orientation (thinking the future market will be "good" or "bad") on the future stock market is the indicator to reflect the investor emotion. In this paper, we proposed a method of sentiment classification and apply it to perform sentiment classification on Sina micro-blog (currently, the largest Chinese micro-blog platform). In detail, our approach contains two-step classifier. Firstly, the first classifier will identify the sentence that contains the future sentiment. Secondly, use the second classifier to classify the sentence which identified by the first classifier into positive or negative. The experimental results show that our method achieves a decent performance on identifying the future stock investor sentiment orientation.

**Keywords** Computer application, Chinese information processing, Investor sentiment, Micro-blog, Sentiment classification, Sentiment orientation

## 1 引言

微博是 Web2.0 时代新兴起的一种集成化、开放化的互联网社交服务。由于微博格式自由、使用方便，越来越多的用户开始从传统的通讯工具（博客、邮件等）转向微博服务。用户可以通过微博发布自己当前的心情、状态，讨论各种各样的话题，对当前的热点问题发表自己的看法。随着微博用户的迅速增长，微博的发布量也在急速增长。在这些海量的文本信息中，有很大一部分是带有情感的文本信息。这些情感文本信息是非常宝贵的意见资源，我们可以利用这些文本信息进行情感文本分类研究。

情感文本分类是按照文本表达的情感倾向性对文本进行分类[1]。例如，判断文本对某个事物的评论是“好”还是“坏”。虽然，该任务研究历史不长，但是其已经成为自然语言处理方向里面的一个研究热点。目前，情感分类任务的研究主要是针对某个事物的评论，而且评论的对象一般都是当前非常具体的静态事物，例如产品、电影、酒店等。而针对未来事物的评论却很少有学者去研究，例如，评论某位候选人是否可以当选、评论未来股市上涨还是下跌、评论某部还未上映的电影值得看还是不值得看等，这些都是对未来事物的评论。本文将

本文受国家自然科学基金项目（61003155，90920004）资助。

庞 磊（1985－），男，硕士生，主要研究方向为自然语言处理；李寿山（1980－），男，副教授，主要研究方向为自然语言处理；张 慧（1985－），女，硕士生，主要研究方向为自然语言处理；周国栋（1967－），男，教授，博士生导师，主要研究方向为自然语言处理。

对微博上股票投资者对股票市场未来行情的情感倾向性识别展开研究。

股票投资者发布的微博中有很大一部分反映了股票投资者的情绪，投资者情绪是研究经济市场走向的重要指标。经济学上，行为金融学认为股票投资者情绪影响投资者决策，进而影响股票市场。而股票投资者情绪都是从投资者对未来股票市场行情的情感倾向上反映出来的，Solt 等人[2]把看涨情绪指标(Bullish Sentiment Index)和看跌情绪指标(Bearish Sentiment Index)都用 BSI 来表示，其用来反映投资者情绪指标。刘超等人[3]通过使用 BSI 情绪指标对投资者情绪和上证综指关系的研究表明，投资者情绪和上证综指有非常直观的联系和比较一致的运行趋势，有很强的相关性。因此，股票投资者对未来股票市场行情的情感倾向与股票市场的走向也有较强的相关性。

我们将对微博上股票投资者对股票市场未来行情的情感倾向性识别展开研究。本文的贡献主要有以下两点：

1)本文首次提出针对未来事物评论进行情感分类研究，并对股票投资者对未来股票市场行情的情感分类进行探索性研究；

2)本文运用自然语言处理技术，自动识别微博上股票投资者对未来股票市场行情的情感倾向性，从而为经济学方面的研究学者提供 BSI 数据。

本文第 2 节对近年在投资者情绪和微博文本情感分类研究上的相关工作进行介绍；第 3 节详细介绍股票投资者未来情感倾向性自动识别方法；第 4 节是实验结果与分析；最后是本文的结论和下一步工作。

## 2 相关工作

### 2.1 投资者情绪

投资者情绪理论是行为金融学的重要组成部分，对于投资者情绪的定义，目前还没有统一的标准。Brown 等人[4]认为投资者情绪代表了市场参与者对某一标准的预期，这个标准就是看涨(或看跌)的投资者的期望收益会高于(或低于)市场的平均收益，即使这一平均收益值无法得到。Baker[5]却认为，投资者情绪可能是一种市场投机倾向，情绪会驱动投机性行为的产生，进而导致股票收益的截面效应，即使市场套利力量在股票间是相同的。Mehra 等人[6]根据行为资产定价理论，认为投资者情绪反应的是投资者对未来股价波动的主观性偏好，尤其反映在风险偏好上。

近年来，国外有少量关于股票网络在线评论情感分析研究的文章，国内还没有这方面的研究。Das 等人[7]针对 Yahoo 网站股票评论进行了研究，文章利用基本词典和赋权值的股票术语词典，使用 5 种文本分类算法并结合投票算法，将评论分为看好(正面评论)、看跌(负面评论)、看平(中立评论) 3 类，并证明设计的情绪指数与股市指数之间存在着有效联系。

### 2.2 微博文本情感分类

情感分类是按照文本表达的情感倾向性对文本进行分类。根据目标载体的粒度可以分为 3 类：篇章级别的情感分类、句子级别的情感分类和词/短语级别的情感分类。Pang[1]首次将机器学习的方法应用于篇章级的情感分类任务中，并指出这种方法比基于规则的分类方法在分类性能上有明显的优势。文献[8-10]在有效特征的发现以及特征选择和特征融合等方面做了相应研究。文献[11]在分类器的选择上和分类器融合等方面做了相应研究。基于特征的监督分类方法是目前主流的情感分类方法。

微博是一种新兴的社交网络服务，目前对微博的情感分类的研究还相对较少。Go 等人[12]利用 Twitter 上的表情符号作为标签收集英文语料作为训练集进行情感分类，省去了人工标注语料的过程；Pak 等人[13]用了文献[12]的方法对英文微博语料进行收集与自动标注，进行情感分析与意见挖掘的研究；Jiang 等人[14]提出在微博上通过加入评价对象相关的特征来提高情感分类的效果；另外，Davidov 等人[15]利用 Twitter 上的标签和笑脸表情符对 Twitter 上的微博语料进行强化学习研究。

股票投资者情绪是一个很难度量的概念，本文将采用文献[4]的观点：以看涨或看跌的评论作为股票投资者情绪指标，利用自然语言处理技术，自动识别微博上股票投资者对未来股票市场行情的情感倾向性(未来市场行情“好”或“坏”)。本文研究的目的不是对股票指数进行预测，而是如何更好地自动识别出看涨或看跌的评论，从而为经济学方面的研究学者提供 BSI 数据。

## 3 股票投资者未来情感倾向性自动识别方法

微博评论文本一般内容短小精悍、口语化、语法不规则，但是要比长篇评论观点明确，包含着更多的微博作者情感。本文的研究目的是自动识别股票投资者对未来市场行情的情感倾向性。在股票评论的微博中，针对未来股票市场行情的评论往往会包含当前市场行情的评论，微博作者为了使自己发表的评论具有说服力，很多情况下根据当前的市场行情来发表对未来市场行情的评论。例如，“今天上证指数收于 2688.75，跌了 82.04，今天大盘的下跌是一次很好的洗盘，也是一个不理性的下跌。从技术上看，明天有望反弹。”这条微博中，第一句是对当前股票市场行情的评论，第二句是根据第一句对未来股票市场行情做出的评论。对新浪微博股票投资者评论的统计显示，在股票投资者对未来股票市场行情的评论中，对当前股票市场行情评论的比例占 68.5%。这样显然会给情感分类任务带来困难，因此，本文的研究任务是基于句子级别的情感分类任务，任务分为两级：

1)识别含有未来股票市场行情评论的句子。

2)对第一级识别出来的句子进行情感分类，分为看涨或看跌两类。

图 1 给出了总体流程，各步骤主要技术内容如下：

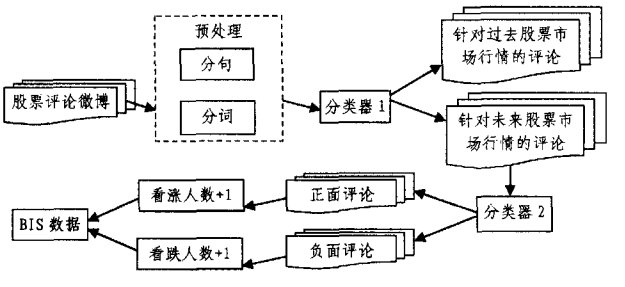


图 1 投资者未来情感倾向性自动识别流程图

在图 1 所示的流程图中，分类器 1 是从时间序列上将文本分为针对过去的评论和针对未来的评论两类；分类器 2 是

将分类器1识别出来的未来评论分为正面评论(看涨)和负面评论(看跌)。BIS数据是根据经济学上的定义进行统计的,其定义如下:

$$BSI(1)=\frac{\text{看涨投资者人数}}{\text{看涨投资者人数}+\text{看跌投资者人数}}\times 100\% \quad (1)$$

$$BSI(2)=\frac{\text{看跌投资者人数}}{\text{看涨投资者人数}+\text{看跌投资者人数}}\times 100\% \quad (2)$$

式中,BSI(1)是看涨情绪指标(Bullish Sentiment Index),BSI(2)是看跌情绪指标(Bearish Sentiment Index)。

分类器1(时间分类器)和分类器2(极性分类器)的准确率直接决定着BSI的准确率,因此本文的研究重点是对两级分类器的构建。

## 4 实验结果及分析

### 4.1 语料收集与标注

本文是首次探索性地提出对未来情感倾向性文本进行分类的研究,所以使用的语料都是人工标注的。从新浪微博上收集了共1700篇股票微博评论,并将这些微博评论进行分句,共生成2756个句子。

通过人工标注将语料分为3类:针对过去的评论、针对未来的正面评论、针对未来的负面评论。为了使人工标注语料达到非常高的准确率,所有语料由两人分别标注一遍,然后将两人标注一致的语料选取为最终语料。其中,针对过去的评论有719篇,针对未来的正面评论有423篇,针对未来的负面评论有458篇。人工标注的部分样例如表1—表3所列。

表1 针对过去的评论

| |
|---|
| 1. 今天,大盘下跌18点,报收于2515点,得一小阴线。 |
| 2. 沪指试图上攻10日线,但是由于上方压力较大,大盘震荡回落。 |
| 3. 在大盘大幅下挫的情况下,主力尾盘逆势拉升更容易引来跟风盘。 |
| 4. 受外围股市重挫消息冲击,今日大盘主动回调并缩量震荡,短期获利盘出现回吐,部分医药股继续走强。 |
| 5. 周二受外盘暴涨刺激,直接以反叛线出手,试图构筑圆弧底右半球。 |

表2 针对未来的正面评论

| |
|---|
| 1. 技术上未走稳,后市看好不变,今天跌,明天可止跌,周3会止跌上扬,判断周4周5为升势。 |
| 2. 今天振幅加宽,市场即将选择突破方向,美债危机缓解,将有利于市场向上突破,上涨第一目标位为十日线。 |
| 3. 本周大盘将收小阳线(指周线,也不排除小阴线,即可能收在2770点附近),下周大盘可能创出本轮反弹新高。 |
| 4. 明日大盘仍要看今晚美国股市脸色,预计该小幅反弹了! |
| 5. 未来依然存在一定的止跌反弹预期。 |

表3 针对未来的负面评论

| |
|---|
| 1. 更可恨的是,大盘惯性走低考验2545点位的概率很大。 |
| 2. 维持下周继续看空大盘,并假设601398下周开始砸盘。 |
| 3. 后市大盘仍将谨慎震荡,市场悲观情绪将使大盘短期难改弱势格局。 |
| 4. 受利空影响,大盘中阴收盘,主动性及主力均为全面净流出,后市将继续振荡下行。 |
| 5. 看期指的盘前走势,估计大盘今天好不到哪去。 |

### 4.2 实验设置

在获得标注语料后,使用机器学习的方法来构建分类器。分类任务是将评论按两级进行分类:第一级是将评论按时间分为针对过去的评论和针对未来的评论;第二级是将针对未来评论分为正面评论和负面评论。本文采用分类正确率来评价分类的效果,其定义如下:

$$ACC=\frac{\text{正确分类文本数}}{\text{测试集中文本总数}}\times 100\% \quad (3)$$

实验中比较了3种分类器,其分别是支持向量机(SVM)、朴素贝叶斯(NB)和最大熵(ME),其中SVM使用的是标准工具light-SVM[1],NB和ME使用的是MALLET机器学习工具包[2]。在使用这些工具的时候,所有参数都设置为默认值。在分类之前,首先采用中国科学院计算机研究所的分词软件ICTCLAS[3]对中文文本进行分词操作。

### 4.3 第一级分类器性能分析

第一级分类器是时间分类器,选择不同的特征来评价时间分类器的性能。在人工标注语料中,针对过去的评论共有719篇,针对未来的评论共有881篇(正面评论+负面评论),为了使实验数据达到平衡,随机地从两类评论中各抽取出700篇语料。实验中,采用5倍交叉验证的方法。实验结果如表4所列。

表4 第一级分类器在不同特征上的分类结果

| | SVM | ME | NB |
|---|---|---|---|
| Unigram | 89.4% | 89.6% | 85.5% |
| Bigram | 88.0% | 88.5% | 86.4% |
| Unigram+Bigram | 90.3% | 90.7% | 87.7% |

从表4的结果可以看出,3种分类器在3种不同的特征上都表现出了非常好的分类效果,特别是以Uingram+Bigram作为特征时,分类效果明显优于前两种特征。从3种分类器的分类效果上来看,ME分类器在分类性能上要优于其他两种分类器;而NB分类器在分类性能上要明显低于其他两种分类器,这可能是由于本文是基于句子级别的分类任务。

针对未来的评论在文本特征表示上与主客观文本分类中的主观文本非常类似,Kim对主观文本的定义为:主观文本要包含4要素〈主题、持有者、陈述,情感〉[16]。情感是区分主客观文本的主要特征,一般地,只有主观文本才含有情感词汇,在客观文本中不会有情感词汇。在时间分类中,也只有针对未来的评论中含有未来时间词汇,在针对过去的评论中不会含有未来时间词汇。相对于情感词汇来说,未来时间词汇要少得多,表5给出了部分未来时间词汇样例。

表5 股票评论中未来时间词汇

| 动词 | 名词 | 副词 | 时间词 |
|---|---|---|---|
| 预计 | 后市 | 必定 | 明天 |
| 估计 | 概率 | 可能 | 下周 |
| 会 | 后续 | 势必 | 下个月 |
| 看涨 | 趋势 | 或许 | 下半年 |
| 看跌 | 短期 | 也许 | 明年 |
| 预期 | 长期 | 大致 | 马上 |
| 有望 | 机会 | 一定 | 即将 |

### 4.4 第二级分类器性能分析

第二级分类器是极性分类器,与第一级分类器的实验相同,我们选择了不同的特征来评价极性分类器的性能。为了

1) http://www.svmlight.joachims.org/

2) http://mallet.cs.umass.edu/

3) http://ictclas.org/

使实验数据达到平衡，随机地抽取了正面评论和负面评论各420篇语料，实验中，采用5倍交叉验证的方法。实验结果如表6所列。

表6 第二级分类器在不同特征上的分类结果

| | SVM | ME | NB |
|---|---|---|---|
| Unigram | 83.2% | 84.7% | 82.5% |
| Bigram | 78.8% | 81.4% | 82.3% |
| Unigram+Bigram | 83.6% | 85.7% | 83.5% |

从表6的结果可以看出，3种分类器在3种不同的特征上都表现出了比较好的分类效果，与第一级分类器分类结果类似，当以Uingram+Bigram作为特征时，分类效果明显优于前两种特征。ME分类器在分类性能上也要优于其他两种分类器，SVM分类器在Bigram特征上分类效果较差，NB分类器虽然在Bigram特征上分类效果最好，但在Uingram+Bigram特征上的分类效果要远低于ME分类器。

股票极性分类是一种特殊的情感文本分类，在分类性能上与一般的情感分类类似；而在文本特征表示上，股票评论情感词汇与一般情感词汇有很大不同，表7给出了部分股票评论情感词汇。

表7 股票评论情感词汇

| 正面情感词 | 负面情感词 | 程度词 |
|---|---|---|
| 上涨 | 下跌 | 明显 |
| 看多 | 看空 | 小幅 |
| 上攻 | 低开 | 大幅 |
| 反弹 | 走低 | 相当 |
| 突破 | 利空 | 最 |
| 攀升 | 悲观 | 非常 |
| 乐观 | 萎缩 | 强烈 |
| 阳线 | 割肉 | 更 |

**结束语** 目前，情感分类任务研究主要是针对某一具体事物的评论，这些事物一般都是评论作者以前接触过的。针对未来事物的评论还没有学者进行研究，而其也是非常有研究价值的，例如，电影发行厂商可能会对电影上映前大众的反应非常感兴趣，通过大众的反应可以调整电影营销策略。本文首次提出针对未来事物评论进行情感分类研究，并对股票投资者对未来股票市场行情的情感分类进行探索性研究。

为了获取规模比较大的文本，我们以微博股票投资者评论作为语料进行情感分类研究，收集并标注了一定规模的语料，并使用机器学习的方法训练分类器，实现股票投资者评论的两级分类。实验结果表明，本文提出的方法在自动识别股票投资者未来情感倾向性上表现出不错的效果。

本文首次提出针对未来事物评论进行情感分类研究，由于语料需要人工标注，因此本文仅对股票领域进行了探索性的研究。下一步将扩充语料，并对更多的领域展开研究。

## 参 考 文 献


[1] Pang B, Lee L, Vaithyanathan S. Thumbs up? Sentiment classification using machine learning techniques[C]// Proceeding of the conference on Empirical Methods in Natural Language Processing(EMNLP). 2002

[2] Solt M E, Statman M. How useful is the sentiment index? [J]. Financial Analysts Journal, 1988, 44(5): 45-55

[3] 刘超，韩泽县. 投资者情绪和上证综指关系的实证研究[J]. 北京理工大学学报：社会科学版，2006，8(2)：57-60

[4] Brown G W, Cliff M T. Investor Sentiment and the Near-term Stock Market[J]. Journal of Empirical Finance, 2004, 11(1): 1-27

[5] Baker M, Stein J. Market liquidity as a sentiment indicator[J]. Journal of Financial Markets, 2004(7): 271-299

[6] Mehra R, Sah R. Mood fluctuations, projection bias, and volatility of equity prices[J]. Journal of Economic Dynam ics and Control, 2002, 26(5): 869-887

[7] 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

[8] Cui H, Mittal V, Datar M. Comparative experiments on sentiment classification for online product reviews[C]// Proceeding of the conference on AAAI. 2006

[9] Kim S, Hovy E. Automatic identification of pro and con reasons in online reviews[C]// Proceeding of the conference on of the 49th Annual Meeting of the Association for Computational Linguistics. 2006

[10] Zhao J, Liu K, Wang G. Adding redundant features for CRFs-based sentence sentiment classification[C]// Proceeding of the conference on Empirical Methods in Natural Language Processing(EMNLP). 2008

[11] 李寿山，黄居仁. 基于Stacking组合分类方法的中文情感分类研究[J]. 中文信息学报，2010，24(5)：56-61

[12] Go A, Bhayani R, Huang L. Twitter Sentiment Classification using Distant Supervision[R]. CS224N Project Report. Stanford, 2009

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

[14] Jiang L, Yu M, Zhou M, et al. Target-dependent Twitter Sentiment classification[C]// Proceeding of the conference on of the 49th Annual Meeting of the Association for Computational Linguistics. 2011

[15] Davidov D, Tsur O, Rappoport A. Enhanced sentiment learning using twitter hashtags and smileys[C]// Proceeding of the 23rd international conference on Computational Linguistics (COLING). 2010

[16] Kim S M, Hovy E. Determining the Sentiment of Opinions [C]// Proceeding of the 23rd international conference on Computational Linguistics (COLING). 2004: 1367-1373

[17] 宗成庆. 统计自然语言处理[M]. 北京：清华大学出版社，2008：340-353