基于 spark 文本分类的股吧情绪指标的构建

司徒雪颖
中央财经大学

一、研究背景与研究目的

社交化媒体中，与金融相关的文本呈现出爆发的趋势。伴随着金融市场，尤其是股票市场的走高与走低，社交化媒体，如微信、微博等，大量的文本数据与其相关联。由于我国股市的个人投资者占大部分，因此在股票市场呈现出大幅震荡的同时，投资者的情绪同样呈现出相似的波动。

投资者情绪历来是金融学领域的热点研究问题。投资者的情绪代表着投资者心理对未来市场的预期，反映了投资者在未来一段时间内，可能进行的投资操作。众多的个人投资者情绪共同构成了金融市场上投资者的情绪指标。

目前，对市场投资者情绪的监测，不再简单的依赖于过去传统的问卷调查等方式，观察投资者情绪的手段呈现出多样化、丰富化的情况。比如，股票论坛中包含着相当数量的金融数据与投资者情感倾向信息，它在一定程度上反映了整个金融市场投资者的投资情绪，通过对论坛投资者情绪的分析，对研究金融市场的可能走向具有一定的帮助。利用互联网金融文本信息，对投资者的情绪进行探究，分析大部分投资者的观点，已经成为了众多研究者关注的课题。

本文通过爬虫获得互联网中大量关于上证指数的帖子文本 119203 条，时间集中在 2017 年 10 月-2018 年 1 月，并对帖子文本进行情绪分析，对投资者的投资心态进行考量，文本情绪分类中采用朴素贝叶斯、随机森林等机器学习的分类方法，利用 spark 实现分类过程，之后以标记了情绪倾向的帖子为依据，构建股吧的情绪指数。研究结果表明：本文提出的金融市场情绪指标与金融市场股指涨跌之间确实存在一定的关联，对互联网社交媒体进行舆情监测有一定的意义。

二、股吧情绪指数的构建

1.股吧帖子文本源数据的获取

本文选择东方财富网的股吧中的上证指数吧进行数据的抓取。东方财富网作为我国历史相对较长的网络金融交流平台，其平台沟通氛围相对其他平台更加良

好，数据中的噪音相对更少，数据更加集中在金融问题的讨论上，数据质量相对更高，并不像其他论坛中可能会出现类似于“灌水”的现象。而之所以选取上证指数吧，在于我们的研究对象为整个金融市场的情绪走向，而非单个股票的可能走势。因此我们选取了相对更有代表性、数据内容更多的上证指数吧来进行研究。

在本次抓取中，一共抓取了股吧 1-1500 页的帖子，总计 119203 条数据，数据的主体为文本形式的数据。数据抓取的方式为 python 的 scrapy 框架，抓取的对象包括帖子标题、作者、阅读量、评论量、发帖时间等，具体的数据结构如表 1 所示：

表 1 抓取的数据介绍

变量	含义	单位	备注
Title	帖子标题	无	无
Writer	帖子作者	无	无
Content	帖子内容	无	去除重复性、内容少于 4 个字的帖子
Comment	评论数	条	0~356
Time	发帖时间	年月日	20171009-20180124
url	帖子网址	无	无

抓取的样例如表 2 所示：

表 2 抓取样例

title	writer	content	comment	time	url
目前看，升势还会延续	晓歌升级	盘面看，今天金融股冲高回落，好在其他版块接力上行使得沪市又再创新高；深市也已经摆脱下行趋势，重返上升势头。目前看，升势还会延续，因为现在市场上的赚钱效应越来越明显，大家也都有了持股的信心。	0	2018/1/24	news,szzs,743048179.html

2.股吧帖子文本预处理

由于抓取到的源数据是掺杂着标点，特殊符号，及对文本含义无意义的语助词和语气词的完整中文语句，不能直接被计算机理解，在做分析前需进行文本预处理。文本预处理主要分为分词，删除停用词，删除低频词，文本向量化处理。

（1）分词

文本分词是指将文章或语句中的词语按照一定标准进行划分的过程。相对于英语文本而言，汉语由于文本之间没有天然的分隔，处理其有一定的难度。将较长的语句或文章转化成较短的单词或词组，这一过程即中文分词。本研究中，采用基于统计的分词方法，通过隐马尔可夫（HMM）模型的 Viterbi 算法得到分词结果，具体分词过程是通过 Python 中 jieba 分词包实现。另外，由于金融领域存在众多专有词汇，如果只用 jieba 包默认的词典进行分析，则会无法识别这些专有词汇，因此在 jieba 包添加了自定义词典，添加的词汇一方面来源于股吧文章的信息提取，另一方面来源于搜狗输入法的金融词库。

（2）去除停用词

去除停用词指过滤文本中的特殊字符和对文本含义无意义的词语。例如“的”，“啊”一类的语气语助词，对文本情感倾向判定无意义，却在文本向量表示时由于占据较大比重而对后续分析造成干扰，降低情感分类的准确性。同时，根据分词文本主题不同，停词表需要进行针对性地修改来提高准确性。因此，研究中用到的停词表在《哈工大停用词表》的基础上，根据帖子文本特点进行了修改。

（3）去除低频词

去除停用词后先做词频统计，取前 50 个词频最大且有意义的词汇绘制词云图。从图 1 中可以看出，除了“涨”，“跌”，“反弹”，“调整”有明显感情倾向性的词外，其他超高频词如“市场”，“大盘”，“指数”，“公司”等均无明显情感倾向性。

图 1 抓取文本的词云图

由于存在大量无意义的低频词（出现的频率仅为 1 次的为低频词，有 23388 个，总词典词汇数为）可能会降低分类精度，因此对去除停用词后的文本再删除低频词。

（4）文本向量化

本研究中文本向量化采用 tf-idf，用稀疏方式储存词-文档矩阵。矩阵维度为 t*n，t 代表文本个数，n 代表词语个数。 用词-文档稀疏矩阵直接进行分类是不可取的，维度过高及矩阵过于稀疏将导致分类精度低，因此向量需先降维。矩阵降维采用非负矩阵分解（NMF）的方法，分解后应用于分类算法的文档向量也非负，因此可以用非负矩阵分解（NMF）方法降维。经过 NMF 分解，文档矩阵作为原始词-文档向量的替代应用到分类算法。

3.股吧帖子文本情感分类

经过降维处理后文本向量可用于后续分类处理。要计算股吧情绪指数，需要先人工给 2000 条随机选取的文本打上积极，中性或者消极的标签，之后均采用自适应学习。本文中分类采用机器学习中的朴素贝叶斯的方法和随机森林的方法，学习已标记文本，得到模型后，再对未标记文本做预测，获取最终标签。标签方法为：“1”表示积极；“-1”表示消极。标记示例如表 3 所示：

表 3 文本情绪标注示例

文本	情感标签
听说，明儿，乐视要复牌，吓得我今天股票全清仓了，就怕它带着来一波跳水调整!!	-1
盘面看，今天金融股冲高回落，好在其他版块接力上行使得沪市又再创新高；深市也已经摆脱下行趋势，重返上升势头。目前看，升势还会延续，因为现在市场上的赚钱效应越来越明显，大家也都有了持股的信心。	1
强烈呼吁维护中小投资者权益，实行 T+0 交易！	0

使用人工标记的 2000 条样本模型建立过程中，采用了朴素贝叶斯（Naïve Beyes）和随机森林（Random Forest）分别训练数据，通过交叉验证得到平均准确度，如表 4 所示：

表 4 机器学习方法分类效果

标准\方法	朴素贝叶斯	随机森林（n=1000）
Accuracy	0.723	0.745

Precision	0.836	0.778
Recall	0.346	0.441

可以看出随机森林算法准确度较高。n 为该算法中包含的树的数量（n=1000），因此选用随机森林算法预测标签。

4.股吧情绪指数计算

股吧情绪指数计算方式如下：以 2018 年第一个交易周为例，该类别下，文本总数为 4507 条，其中被标记为“1”的文本共 464 条，标记为“-1”的文本共 3637 条，因此 2018 年第一个交易周的股吧情绪指数为：

$$\frac{464-3637}{464+3637}=-0.7737$$

即计算所有有正向、负向情感倾向的文本的得分均值。这种计算方法忽略了被标记为 0 的大多数文本，有效排除广告等无意义文本在情感指数计算中的影响。从计算方式可以看出，单方面得分将在-1 到 1 间波动，情感指数将在-1 到 1 之间波动。由于爬取的绝大部分帖子是 2017 年 10 月（即 2017 年第 41 周）以后的，只有 2.8%的帖子的发表时间在 2017 年 10 月之前，因此只选择 2017 年第 41 周以后的帖子计算情绪指数。按照上述方法，基于随机森林算法模型计算出 2017 第 41 周至 2018 年第 4 周股吧的情感指数如表 5 所示：

表 5　2017 第 41 周-2018 年第 4 周情绪指数

周序号	情绪指数	帖子数
2017-41	-0.7534	4618
2017-42	-0.7750	2329
2017-43	-0.6974	948
2017-44	-0.7587	1368
2017-45	-0.7394	2107
2017-46	-0.8589	9959
2017-47	-0.8788	11715
2017-48	-0.8571	10230
2017-49	-0.8902	14110
2017-50	-0.8445	8693
2017-51	-0.8507	7606
2017-52	-0.8571	7196
2018-01	-0.7737	4507
2018-02	-0.7334	7463
2018-03	-0.8165	9337
2018-04	-0.8186	4251

从表格可以看出，股民的情绪越低迷，发帖数目越多，帖子数与情绪指数的相关系数为-0.8358，高度的负相关也证明了这一点。

三、实证分析

1. 每日股吧情绪指数

从下图可以看出，股吧的股民情绪一直处于负值，说明股民情绪在熊市是倾向负面的，在 2017 年 11 中下旬情绪达到底部，而后又回升，2018 年 1 月初达高位又回落。

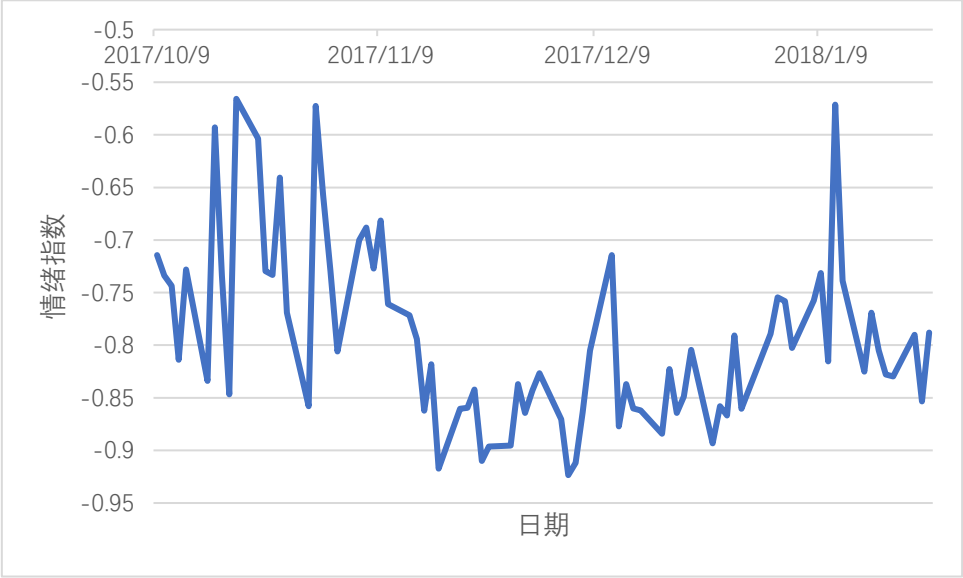

图 2 每日股吧情绪指数

2. 与上证指数比较

去除周末节假日等日期的情绪指数后，将股吧交易日的情绪指数与上证指数相比较。从图中可以看出，情绪指数与上证指数的周均收盘价趋势较为一致，实际上两者的相关性为 0.66，这一结果是可以预见的。因为在人工标记文本时，不同时间段均有大部分的股民发帖称股市出现“指数上涨却千股跌停”等奇观，而股民情绪与他们购买的个股涨跌比大盘涨跌更为相关。另外，市场小幅波动总能造成股民们较大的情绪波动，这一点在图中 10 月指数持平而情绪指数剧烈震荡，11 月股民情绪比指数先一步下降至最低点，两处得以体现。

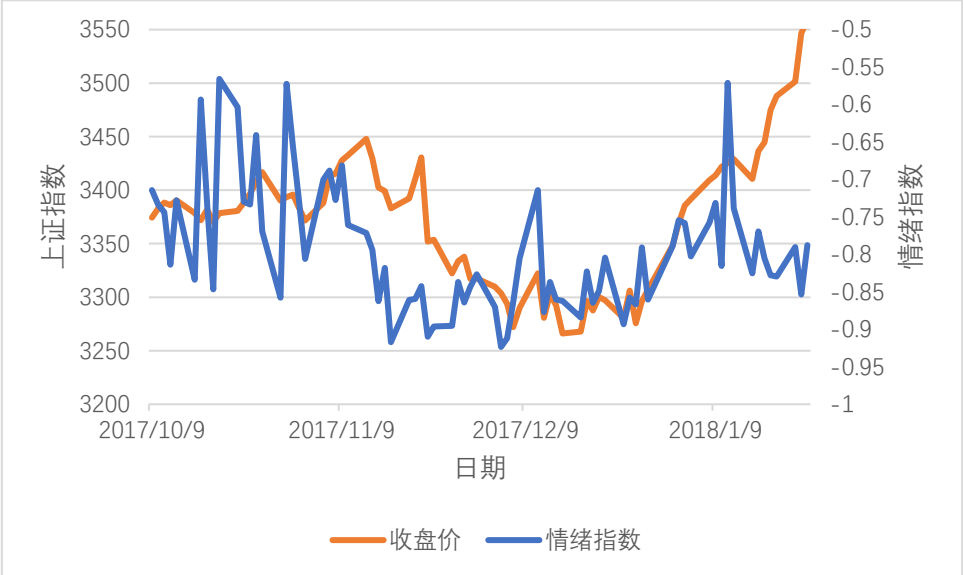

图 3 股吧情绪指数与上证指数的比较

3.总结

通过互联网文本数据构建股吧情绪指数作为预测股民信心指数的变量，利用了互联网信息集中和时效性强的特点，减少传统消费者信心指数调查工作量，节约人力、物力成本，并使股吧情绪指数指数更加有效和准确。

四、结论与不足

本文通过文本情感分析和机器学习的方法，衡量了股吧这一典型的网络社交媒体的舆论情感倾向，以文本的形式，切入了金融市场情绪的检测，为股吧等一类典型的社交媒体监测舆情提供了一种方式。

此外，本研究存在一些局限性。首先数据获取自 2017 年 10 月开始，整体尚不足一年，一定程度上使得指数说服力存疑。另外文本分类时精度不够高，存在一定的误分文本，造成股吧情绪衡量有一定误差，希望在更进一步的研究中进行改进。