运用新闻和公司信息预测医疗行业股市波动

钱丽玥
SIST
ShanghaiTech University
Shanghai, China
qianly@shanghaitech.edu.cn

沈昭正
SIST
ShanghaiTech University
Shanghai, China
shenzhzh@shanghaitech.edu.cn

井文可
SIST
ShanghaiTech University
Shanghai, China
jingwk@shanghaitech.edu.cn

1 INTRODUCTION

金融科技高速发展的今天，股票市场已经成为了一个国家金融体系中非常重要的一部分。长期以来，由于股票价格具有高度波动性，使得股票价格的预测成了分析师和研究人员关注的话题。传统股票价格的预测大多局限于银行板块数据系列，股票收盘价，月度加权股值等 [1]。

但是最近的研究表明，公共领域的大量在线信息，如主流媒体的新闻，社交媒体讨论，政府发布的官方文案都可能会对投资者对金融市场的看法产生很大的影响 [3]。在自媒体高速发展的今天这一点更加明显。财经新闻以一种公开的信息披露方式，对我国股市产生了非常重要的影响，其中包括政策扶持类财经新闻、兼收并购类财经新闻、再融资类财经新闻、盈利类财经新闻、违规处罚类财经新闻。尽管它们对股市的影响不尽相同，但在我们的课程作业中将不区分新闻本身的属性，而只关注新闻对于股市的利好利空影响。目前已经有包括支持向量机 [3] ，贝叶斯网络 [4]和深度学习 [5]等算法在各个级别上进行了几种分析研究，结果表明股票价格的变动与新闻文章的发布之间存在很强的相关性。

同时近些年来，随着人民生活水平的提高和对医疗保健需求的不断增长，我国医药行业一直保持着较快的增长速度，在国家政府对行业的大力扶持下，医药行业逐渐成为国民经济中发展最快的行业之一。2020 年世界范围内新冠肺炎疫情的爆发，进一步提高了国家乃至世界对医药行业的重视程度，并且加大了在资金、技术、政策方面的支持和帮助。此外，在人口老龄化以及中国城镇化加速的背景下，居民收入普遍增长加上人们愈发重视医疗保健，医保体系进一步健全，这些都将刺激并提升中国医药消费需求，推动整个医药行业进行新的变革并形成新格局。因此在我们的课题研究中，将着重探讨医药行业的新闻对股票价格的影响。

2 RELATED WORK

我们的课题研究总体可以分为两个部分，新闻情绪分析以及股票预测。情绪分析作为一种 NLP 技术，主要用于挖掘和评估文本/演讲中表达的意见 [6]，目前针对情绪分析的研究也越来越多，主要是看中其潜在的应用价值 [7]。我们这里主要针对新闻中包含的情绪来预测对股票价格的影响。小部分文章理解新闻文章背后的情感是直接将整条新闻输入神经网络进行训练，大多数的方法是识别新闻文章中的重要词语及其极性。如考虑到一个单词可能既可能在利好新闻中出现，也可能在利空新闻中出现，M. I.Yasef Kaya 等人选择以“名词+动词”这种短语的形式输入网络训练得到单词极性 [8]。张泽亚等人则是提出了一种基于单词点互信息的新闻特征抽取方法，首先选取利好关键词和利空关键词的种子集合，然后使用最优化方法选出利好与利空的标准关键词集合，并用标准关键词集合计算其他单词的利好极性 [9]。Saloni Mohan 等人利用 NLTK 库来分析新闻文本的情感极性，包括积极和消极情感(N, P)。每个新闻文本中的极性标识为最大绝对极性，即 Polarity = (+/−)max(abs(N, P))，而一个公司的最终极性等于其对应的所有文本的平均极性 [10]。而梁士利 [11]等人利用 jieba 分词工具来统计每个词出现的频率，人工筛选标注出不同情感极性的词语，建立专属的金融情感词典。

在股票预测方面，Saloni Mohan 等人将当前新闻的极性和时间 t 之前的股票价格 Price_(t-1)来训练 RNN 模型 [10]。贝勒大学的赵磊 [12]提出了一种离群值挖掘算法，根据股市高频滴答声数据的数量序列来检测异常，使用交易量分布异常来预测股票价格的上升趋势。

3 AIMS AND OBJECTIVES

新闻作为信息传播的一大媒体，它对股市波动有着强烈的影响。例如论文 [13]中就有加入新闻与不加入新闻输入的模型对比，说明了新闻的重要性。一则新闻的发布，对新闻所关注的公司有着或轻或重的影响，影响程度与新闻的发布时间、发布平台、正文内容、点击量、标题等有关。我们主要关注标题的发布时间、内容对股市波动的影响。

同时，不同新闻的点击量与新闻发布的平台相关，点击量越多的新闻影响力也越大，对股票波动的程度影响也越大。所以，我们打算同时将新闻发布的平台这一因素纳入对股票波动预测的分析当中。

总的来说，我们的目标包括下面几个方向：

1. 通过新闻文本的关键信息提取对所选股票的涨跌幅预测
2. 考虑新闻点击量对股票波动幅度的影响

4 PROGRESS

4.1 Preprocessing

我们的处理过程如图 1 所示，

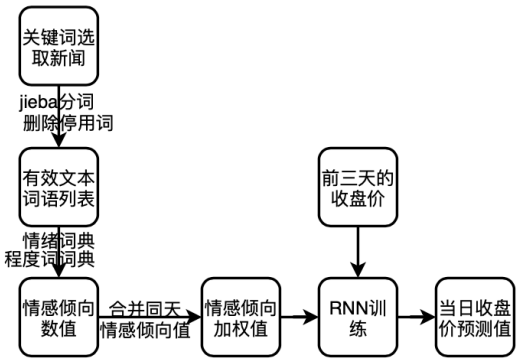

图 1 预处理、训练流程

首先，我们从东方财富网网页上爬取到我们所研究股票从 2020 年 5 月 25 日到 2021 年 5 月 23 日的新闻，其中包括热帖和资讯。如图 2 所示，图 2(a)表示的是复星医药的热帖信息，图 2(b)表示的是复星医药的资讯信息。这些信息的内容有新闻阅读量、评论量、标题、作者以及发布时间。

阅读	评论	标题	作者	最后更新
66949	555	复星医药，这一波底部极限在37-39元（39.01，37.50区	何然静	05-24 15:36
14096	36	新冠疫苗获批在即 复星医药"赶进度"：工厂已落地 3.	寻100倍股	05-24 15:17
31853	228	朋友们，下午好。这次再次压中复星医药阶段性高点和低	何然静	05-24 13:13

(a)

阅读	评论	标题	作者	发帖时间
2656	1	复星医药本周融资净买入1.26亿元，居医药制造板块第九	两融追踪	05-23 15:05
22508	135	复星医药：愿意将疫苗服务于台湾同胞	复星医药资讯	05-22 16:09
637	0	复星医药(600196)融资融券信息(05-21)	复星医药资讯	05-22 07:40

(b)

图 2: (a)热帖信息; (b)资讯信息

我们一共获取了三个医药公司的新闻信息，他们分别是复星医药、沃森生物和恒瑞医药。他们的热帖和资讯新闻数量如表 1 所示。

表 1 新闻数量

公司名称 \ 新闻种类	热帖	资讯
复星医药	946	4471
沃森生物	1048	2511
恒瑞医药	286	3603

然后，我们利用 jieba 分词工具对新闻标题进行分词处理，由于新闻标题中存在很多与情感倾向无关的词语，所以我们再根据哈工大停用词词典去除掉文本中的停用词。

接下来，我们自己建立了与医药行业财经新闻相关的积极情感词典与消极情感词典，找到新闻标题中的积极词和消极词分析情感倾向。不同于 Yasef Kaya 等人选择将新闻分为一个个“名词+动词”短语输入网络训练 [8]，我们选择了一个现有的 NTUSD-Fin 词典，该词典已经根据大量的新闻得到了每一个单词的最终值，可以理解为每一个词语积极消极的抵消值 [14]。我们选取 NTUSD-Fin 中词语极性值大于 0 的值为积极词语，小于 0 的词语为消极词语。再根据我们研究新闻的行业特点建立了我们的情感词典，并结合知网情感分析用词语集 beta 版 [15]中的程度词词典和否定词词典得到了我们使用的总情感词典，具体内容介绍如表 2 所示。

表 2 情感词典信息

情感极性	数量/个	示例
most	69	非常、极端
very	42	格外、很
more	37	较、还要
ish	29	略微、一些
insufficiently	12	轻度、不怎么
pos	162	上涨、增加
neg	421	下滑、损失
inverse	58	不、没有

我们利用该情感词典，通过其中的程度词词典和否定词词典对积极词和消极词加权，最终得到整篇新闻的积极倾向值与消极倾向值。最后把同一天中的新闻合并，根据新闻的

阅读量加权得到这一天中该只股票的积极倾向加权值与消极倾向加权值。

由于我们依赖的是情感分析，所以我们验证了情感分析的准确程度。在图 3 中，绿色线表示的是每一天股价的收盘价，蓝线表示的是每天情感倾向的绝对值，即每天得到的积极倾向值减去消极倾向值的绝对值。可以看到，在图中圆形标出的几个地方，两条曲线的都有同样的峰值，而总体趋势也基本相同，所以说我们的情感分析还是基本可靠的。

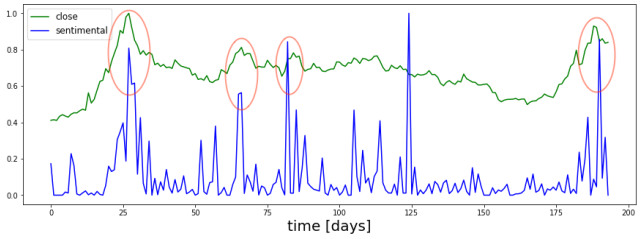

图 3 情感分析的准确性

4.2 Training

对于某只药业股票，我们从聚宽量化交易平台上提取了其从 2020 年 5 月 25 日至 2021 年 5 月 23 日的股价，原有股票数据有 7 个特征（包括日期 date、开盘价 open、收盘价 close 等等），如表 3 所示；得到的药业相关的新闻数据有 3 个特征（日期 date、正向情感程度 pos、负向情感程度 neg）。首先我们将处理好的新闻数据和股票数据按照时间整合起来，得到情感分析与新闻数据结合数据集，其包含包括日期 date、收盘价 close、正向情感程度 pos、负向情感程度 neg 等 9 个特征，结果如表 4 所示。

表 3 股票数据

	date	open	close	high	low	volume	money
0	2020-05-25	30.91	31.12	31.52	30.78	20599861.0	6.412297e+08
1	2020-05-26	31.33	31.33	31.66	31.16	18963313.0	5.943644e+08
2	2020-05-27	31.37	30.80	31.47	30.37	27555024.0	8.486539e+08
3	2020-05-28	30.68	30.11	31.03	29.75	30655938.0	9.267066e+08
4	2020-05-29	30.15	30.54	30.74	30.12	23890232.0	7.284473e+08

表 4 股票数据和新闻数据整合

	date	pos	neg	open	close	high	low	volume	money
0	2020-05-25	0.451054	0.025495	30.91	31.12	31.52	30.78	20599861.0	6.412297e+08
1	2020-05-26	0.052651	0.193771	31.33	31.33	31.66	31.16	18963313.0	5.943644e+08
2	2020-05-27	0.239771	0.046422	31.37	30.80	31.47	30.37	27555024.0	8.486539e+08
3	2020-05-28	0.092130	0.100642	30.68	30.11	31.03	29.75	30655938.0	9.267066e+08
4	2020-05-29	0.094521	0.053147	30.15	30.54	30.74	30.12	23890232.0	7.284473e+08

接着，我们对合成数据集进行了简单的可视化分析。我们分析了股票的基本信息随日期的变化，图 4 为股票数据特征（左图为开盘价、收盘价、最高价、最低价，右图为交易量）随天数的变化趋势。合成数据集的各特征分布如图 5 所示。

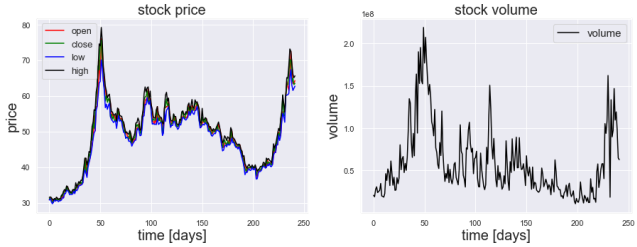

图 4 合成数据集的可视化分析

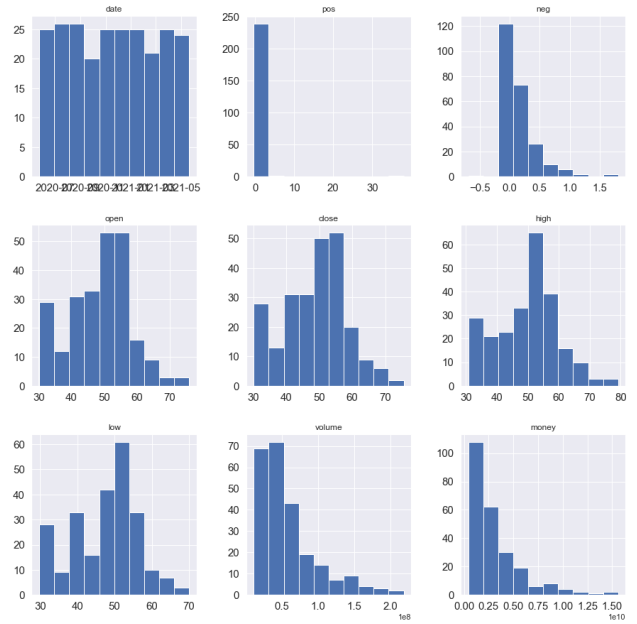

图 5 合成数据集的各特征分布

我们将新的数据集绘制出一个相关性矩阵，以查看各特征之间的相关性。图 6 中可以看出，股票特征之间相关性相对较强（open、high 相关性接近 1，open、volume 相关性在 0.5 以上），而情感特征和股票特征之间的相关性较弱（neg 和 open 相关性接近 0.2）。由于 pos、neg、close 的变量相关性较弱，它们之间不存在重复特征，在之后的神经网络训练中可以将 pos、high、close 筛选出来都放入训练。

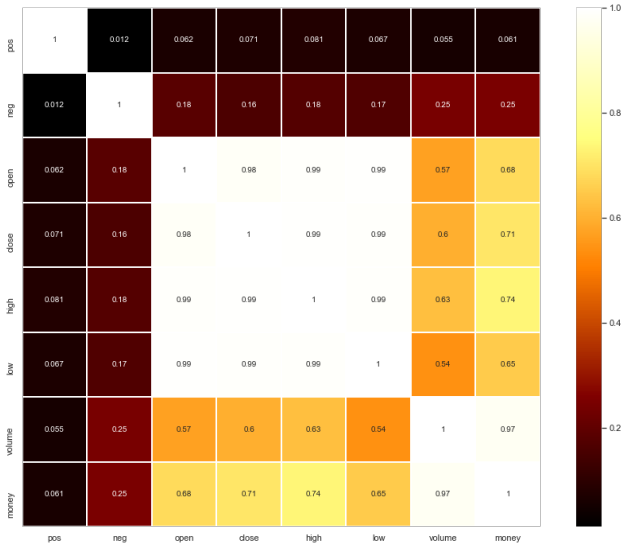

图 6 相关性矩阵

为了对比新闻是否对预测结果造成了影响，以及不同的新闻类型（如官方的资讯新闻以及非官方的股吧论坛）是否对预测结果造成了不同程度的影响，我们设置了不同的对比实验。另外，因为股价数据与新闻数据在时间上不是完全匹配的，即周末有新闻但是没有股价，我们将这几天的新闻去除。对于有些工作日有股价但是没有新闻数据的情况，我们对比了将当天缺失的新闻数据填补为前一天的新闻数据和直接删除无新闻那天数据的结果。此外，我们还对比了利用前一天新闻预测当天股价、利用当天新闻预测当天股价对预测结果的影响。

$c_{t-l}, c_{t-l+1}, \ldots, c_{t-1}, p, n$数据作为输入，来预测输出$c_t$。$c_t$为 t 时间的收盘价，$p$为$t-1$天或$t$天的新闻正向情感程度，$n$为$t-1$天或$t$天的新闻负向情感程度，l 为滞后参数。模型通过获取 t 天以前 l 天的股票收盘价数据，以及$t-1$天或t天的新闻情感，来预测 t 天的收盘价。我们设计的 LSTM 网络如图，包括 30651 个参数。我们通过训练 50 个 epoch，滞后参数为 3，设置前 80%数据作为训练数据，后 20%数据作为测试数据。我们利用准确率(ACC)和均方误差(MSE)作为评估指标对测试结果进行分析，ACC 为股票涨跌情况的准确率，它反映了预测趋势的优劣；MSE 为预测值与真实值之间的误差，计算公式为$\frac{1}{n}\sum_i(p_i - c_i)^2$，它反映了预测数值的优劣。

5 RESULTS

最后的实验结果如图 7，可见我们的预测结果在趋势上与测试数据基本相同。

图 7 股票收盘价预测结果：横轴表示时间，纵轴表示股价。蓝色线表示训练数据集，绿色线表示测试数据预测结果，黄色线表示测试数据。

考虑到新闻情感对实验结果的影响，我们分别进行了仅提取股市收盘价信息，和提取新闻情感及股市收盘价对股票涨跌的影响的实验对比，统计了测试数据集上的 ACC 和 MSE。

首先我们仅将股票数据与新闻数据按照时间取内积，即只考虑工作日时有新闻数据情况的股票预测，预测结果如表 5 和表 6 所示。从两张表的对比中我们可以看出来，使用股票数据与新闻数据内积数据时，使用当天新闻预测当天股价的 ACC 普遍上优于使用前一天的新闻预测当天股价和不使用新闻，然而使用前一天新闻的 MSE 普遍上优于使用当天的新闻和不使用新闻的情况。这表示了使用当天新闻对预测趋势更有效，而使用前一天的新闻对预测价格更有效。

表 5 股票数据与新闻数据内积的 ACC

ACC	新闻类型	用第i-1天新闻	用第i天新闻	不加新闻
复星医药	热点	55.263%	**57.895%**	52.632%
	资讯	**54.348%**	50.000%	47.826%
	热点+资讯	52.174%	**54.348%**	47.826%
沃森生物	热点	45.833%	**53.846%**	51.282%
	资讯	**44.681%**	**44.681%**	**44.681%**
	热点+资讯	**44.681%**	**44.681%**	**44.681%**
恒瑞医药	热点	39.130%	**43.478%**	**43.478%**
	资讯	44.681%	**46.809%**	44.681%
	热点+资讯	44.681%	**48.936%**	44.681%

表 6 股票数据与新闻数据内积的 MSE

MSE	新闻类型	用第i-1天新闻	用第i天新闻	不加新闻
复星医药	热点	**8.011**	8.221	11.356
	资讯	6.859	**6.698**	7.723
	热点+资讯	**6.383**	6.983	7.723
沃森生物	热点	**5.602**	10.002	8.733
	资讯	**4.888**	5.715	5.628
	热点+资讯	**5.104**	5.447	5.628
恒瑞医药	热点	17.308	**14.998**	18.905
	资讯	**5.665**	5.745	5.668
	热点+资讯	5.651	**5.370**	5.668

接着我们将无新闻但是有股价的情况下，当天缺失的新闻数据填补为前一天的新闻数据，预测结果如表 7 和表 8 所示。从两张表的对比中我们可以看出来，使用前一天的新闻数据填补当天缺失的新闻数据时，不使用新闻的 ACC 普遍上

优于使用新闻，使用当天新闻预测当天股价的 MSE 普遍上优于使用前一天的新闻和不使用新闻。对于这种结果，我们认为在填补为前一天新闻数据的情况下，数据本身的情感倾向可能与实际不符合，所以会导致 ACC 在家了新闻之后产生了适得其反的效果；而 MSE 表明了使用新闻还是比不使用新闻的效果更好的。

表 7 填补为前一天的新闻数据的 ACC

ACC	新闻类型	用第i-1天新闻	用第i天新闻	不加新闻
复星医药	热点	41.667%	45.833%	**47.917%**
	资讯	**47.917%**	45.833%	**47.917%**
	热点+资讯	45.833%	45.833%	**47.917%**
沃森生物	热点	46.154%	**47.917%**	**47.917%**
	资讯	41.667%	41.667%	**47.917%**
	热点+资讯	43.750%	43.750%	**47.917%**
恒瑞医药	热点	**50.000%**	43.750%	41.667%
	资讯	**43.750%**	**43.750%**	41.667%
	热点+资讯	43.750%	**45.833%**	41.667%

表 8 填补为前一天的新闻数据的 MSE

MSE	新闻类型	用第i-1天新闻	用第i天新闻	不加新闻
复星医药	热点	5.107	**4.932**	5.826
	资讯	5.029	**4.882**	5.826
	热点+资讯	**4.520**	5.004	5.826
沃森生物	热点	9.905	**5.037**	6.682
	资讯	**5.348**	6.024	6.682
	热点+资讯	5.077	**4.985**	6.682
恒瑞医药	热点	**3.562**	4.121	4.521
	资讯	3.354	**3.192**	4.521
	热点+资讯	3.435	**3.161**	4.521

6 ISSUES AND CHALLENGES

1. 采用的 jiaba 分词器不够完备，例如不能将双重否定词组（例如“难道不是”）保留，而是分成了“难道”与“不是”，导致最后的情感分析出错。

2. 情感分析词库不够全面，有待扩充和修正。

3. 根据公司的新闻去预测可能存在的股票涨跌情况，但是事实上可能存在新闻标题是反向预测股票真正的走势，这会给预测结果带来一些偏差。

4. 同时也存在有些股市新闻标题中并不包含任何利好或者利空的关键词，对股票的预测带来了一定难度。

5. 新闻的渠道挖掘的还不全面，仅考虑了股吧中人们的反应和资讯新闻，短视频等新兴渠道还未涉及。

7 TIMETABLE AND WORKLOAD DIVISION

任务列表	week1-4	Week5-8	week9-10	Week11-14
阅读文献	沈昭正、井文可、钱丽玥			
讨论实现方法	沈昭正、井文可、钱丽玥			
爬取股票数据及预处理		沈昭正		
筛选新闻数据及预处理		井文可、钱丽玥		
训练		沈昭正、井文可、钱丽玥		
测试			沈昭正、井文可、钱丽玥	
分析实验结果			沈昭正、井文可、钱丽玥	
完成报告、ppt				沈昭正、井文可、钱丽玥

图 8: 任务列表的分工和完成情况

REFERENCES

[1] E. Guresen, G. Kayakutlu, and T. U. Daim, "Using artificial neural network models in stock market index prediction," Expert Systems with Applications, vol. 38, no. 8, pp. 10389-10397, 2011.Conference Short Name:WOODSTOCK'18

[2] D. Shah, H. Isah and F. Zulkernine, "Predicting the Effects of News Sentiments on the Stock Market," 2018 IEEE International Conference on Big Data (Big Data), 2018, pp. 4705-4708, doi: 10.1109/BigData.2018.8621884.

[3] C.-Y. Yeh, C.-W. Huang, and S.-J. Lee, "A multiple-kernel support vector regression approach for stock market price forecasting," Expert Systems with Applications, vol. 38, no. 3, pp. 2177-2186, 2011.

[4] L. S. Malagrino, N. T. Roman, and A. M. Monteiro, "Forecasting stock market index daily direction: A Bayesia Network approach," Expert Systems with Applications, vol. 105, pp. 11-22, 2018.

[5] E. Chong, C. Han, and F. C. Park, "Deep learning networks for stock market analysis and prediction: Methodology, data representations, and case studies," Expert Systems with Applications, vol. 83, pp. 187- 205, 2017.

[6] B. Pang and L. Lee, "Opinion mining and sentiment analysis," Foundations and Trends® in Information Retrieval, vol. 2, no. 1–2, pp. 1-135, 2008.

[7] H. Isah, "Social Data Mining for Crime Intelligence: Contributions to Social Data Quality Assessment and Prediction Methods," University of Bradford, 2017.

[8] M.I. Yasef Kaya, Karsl1Gil M E . Stock Price Prediction Using Financial News Articles[C]// 0.

[9] 张泽亚,黄丽明,陈翀,闫宏飞.基于新闻特征抽取和循环神经网络的股票预测方法[J].文献与数据学报,2020,2(01):45-56

[10] S. Mohan, S. Mullapudi, S. Sammeta, P. Vijayvergia and D. C. Anastasiu, "Stock Price Prediction Using News Sentiment Analysis," 2019 IEEE Fifth International Conference on Big Data Computing Service and Applications (BigDataService), Newark, CA,USA,2019,pp.205-208,doi:10.1109/BigDataService.2019.00035.

[11] 梁士利,陈翌昕,陈培培,孙丽敏.基于社交情感数据挖掘的股票市场预测研究 [J].东北师大学报(自然科学版),2020,52(03):105-110.

[12] ZHAO L, WANG L. Price Trend Prediction of Stock Market Using Outlier Data Mining Algorithm; proceedings of the Big Data and Cloud Computing (BDCloud), 2015 IEEE Fifth International Conference on, F, 2015 [C]. IEEE.

[13] Robert P. Schumaker and Hsinchun Chen. 2009. Textual analysis of stock market prediction using breaking financial news: The AZFin text system. ACM Trans. Inf. Syst. 27, 2, Article 12 (February2009),19pages.DOI:https://doi.org/10.1145/1462198.1462204.

[14] Chen C C， Huang H H， Chen H H . NTUSD-Fin: A Market Sentiment Dictionary for Financial Social Media Data Applications.
[15] 董振东.知网[CP/OL]. [2012-03-24]. http://www.keenage.com