分类号_____ 学号 M201877244
学校代码 10487 密级_____

华中科技大学

硕士学位论文

基于新闻情感量化和 LSTM 网络的股票预测模型设计与实现

学位申请人： 廖畅

学 科 专 业： 计算机技术

指 导 教 师： 鲁宏伟 教授

答 辩 日 期： 2020 年 6 月 4 日

A Thesis Submitted in Partial Fulfillment of the Requirements
For the Degree of Master of Engineering

Design and Implementation of Stock Forecasting Model Based on News Sentiment Quantification And LSTM Network

Candidate : Chang Liao

Major : Computer Technology

Supervisor: Prof. Hongwei Lu

Huazhong University of Science and Technology
Wuhan, Hubei 430074, P. R. China

June, 2020

独创性声明

本人声明所呈交的学位论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除文中已经标明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的研究成果。对本文的研究做出贡献的个人和集体，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： 2020 年 6 月 4 日

学位论文版权使用授权书

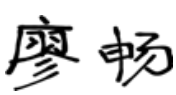
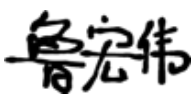
本学位论文作者完全了解学校有关保留、使用学位论文的规定，即：学校有权保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅。本人授权华中科技大学可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存和汇编本学位论文。

保密□，在___年解密后适用本授权书。

本论文属于

不保密☑。

(请在以上方框内打“√”)

学位论文作者签名：  指导老师签名： 

日期： 2020 年 6 月 4 日 日期： 2020 年 6 月 4 日

摘 要

股票市场是资本市场的重要组成部分，企业通过上市融资，可以在市场上获得更多的资金；而投资者则希望通过分析市场上的信息，判断各公司股票的涨跌趋势，进行买入卖出操作从而获取收益。但投资往往都伴随着风险，如果不对影响股票涨跌的因素进行具体分析就盲目地投资，也会给投资者带来一定损失。

为了利用财经新闻信息对股票涨跌进行更准确的预测，进行了股票价格及其相关财经新闻的数据采集方法的研究，使用采集的三只股票的数据进行特征工程，并提取出价格特征和情感量化特征。建立股票的涨跌预测模型，将提取的特征作为输入得出股票价格走势的结果，并且对已有模型进行对比实验。

针对采集股票预测所使用的数据，使用了 Tushare 接口采集股票价格的详细数据，并设计了一种基于 Selenium、Beautiful Soup 工具的股票新闻采集方法，通过模拟浏览器获取新闻网页的过程，定位新闻数据在 HTML（Hyper Text Markup Language）的 DOM（Document Object Model） Tree 中的位置，进而自动化地采集所需股票的新闻数据。

设计了一种基于朴素贝叶斯的新闻情感量化模型，通过该模型可以对新闻文本进行情感分类，并对特定股票每天的所有新闻进行量化，得到股票当天的新闻情感值。实验结果证明该模型对新闻文本情感分类有较高的准确率。

使用 LSTM-Attention 网络，建立股票价格涨跌预测模型，将股票价格特征与新闻情感特征作为模型输入，得出预测的涨跌结果。并且 SVM（Support Vector Machine）、LSTM（Long Short-Term Memory）等模型的预测结果进行对比实验。实现结果证明 LSTM-Attention 网络比传统的 SVM 模型和单一的 LSTM 模型有更好的预测效果，并且加入了新闻情感特征后，模型可取得更高的准确率。

关键词：股票涨跌预测，新闻情感量化，神经网络

Abstract

The stock market is an important part of the capital market, enterprises can get more funds in the market through listing financing, while investors want to analyze the information in the market, judge the up and down trend of each company's stock, and carry out the buying and selling operation to obtain income. However, investment is often accompanied by risks. If we don't make a specific analysis of the factors that affect the rise and fall of stocks, we will blindly invest, which will also bring some losses to investors.

In order to make use of financial news information to predict the rise and fall of stocks more accurately, this paper studies the data collection methods of stock price and its related financial news, uses the data collected from three stocks to carry out feature engineering, and proposes the price features and emotional quantitative features. This paper establishes the stock up and down prediction model, takes the extracted features as the input to get the stock price trend results, and carries on the contrast experiment to the existing models. A kind of

In order to collect the data used in stock forecast, the tushare interface is used to collect the detailed data of stock price, and a method of stock news collection based on selenium and beautiful soup is designed. The process of obtaining news web page by simulating browser is used to locate the news data in DOM (Document Object Model) of HTML (Hyper Text Markup Language) Location in tree, and then automatically collect the news data of the required stocks.

This paper designs a kind of news sentiment quantification model based on Naive Bayes, through which we can classify the sentiment of news text, and quantify all the news of a specific stock every day to get the sentiment value of the stock news on that day. The experimental results show that the model has a high accuracy for emotional classification of news texts.

Using LSTM attention network, the paper establishes a stock price up and down

prediction model, takes stock price characteristics and news emotion characteristics as model inputs, and obtains the predicted up and down results. And the prediction results of SVM (Support Vector Machine), LSTM (Long Short-Term Memory) and other models were compared. The results show that LSTM attention network has better prediction effect than traditional SVM model and single LSTM model, and the model can achieve higher accuracy after adding news emotion features.

Keywords: Stock Trend Prediction, News Sentiment Quantification, Neural Networks

目 录

1 绪论

1.1 研究背景及意义

股票是一种有价证券，股份有限公司将其所有权用这种股份证书的形式进行分配，进而筹得资金。投资者获得股票成为公司股东，并获得公司发展或股价上涨带来的利润，同时也要承担股价下跌带来的风险。所以股票价格的涨跌变化是所有投资者关注研究的重中之重。

股票预测问题随着股票市场的建立而一直存在。Eugene Fama[1]于 1970 年深化并提出了效率市场假说，也称有效市场假说[1]。他认为，股票市场是“信息有效”市场，即股票价格充分反映了已经发生的事件，以及未来可能发生的事件对股票价格的影响，所有的信息已经在当前的股价中体现出来了。所以该假说意味着股价是没有办法预测的。

该假设的成立有三种条件：一是所有投资者都能理性的利用获得的信息获取更高的利益；二是股票市场对新的信息反映迅速而准确，股票价格能完全反映所有信息；三是股票价格变动反映了投资者供求的平衡，即想买的人刚好与想卖的人相等。

然而现实当中要同时满足这三种条件是非常困难的，投资者不一定能随时保持理性，而市场上的信息流通也不一定能那么迅速。许多研究者也证明当今的股票市场并非真正意义上的强式有效市场，所以从技术层面来说，预测股票的涨跌是可能的。

当今时代，随着计算机科学技术的不断进步，在基金公司和证券投资公司中，量化分析正逐渐从人工走向智能。并且随着硬件条件的提升，深度学习技术真正得以普遍性应用，许多领域的工作都因此取得了突破性的进展。前有谷歌 AlphaGo 战胜世界围棋冠军，后有腾讯“觉悟”战胜王者荣耀职业选手，计算机也非常有可能在股票市场上超越证券分析师。

并且股票新闻与股票价格之前有着非常紧密的联系，许多研究者[2-7]都把与股票相关的新闻作为预测股票价格的重要依据。但是，新闻文本本身难以直接用于股价预测，人们需要从新闻中提取有效的市场信息才能发挥新闻的价值。这本质上是一个自然语

言处理的问题，但因为特征抽取是针对金融市场的，所以在自然语言处理室需要结合这一领域的相关知识，这无疑给股票预测问题带来了新的困难。

综上所述，一是股票市场在经济中有着重要的地位，二是股票对于个人和公司都有重要的价值，三是新闻信息与股票的价格之间有着密不可分的关系，四是股票预测在今天正愈加具备可行性。因此，结合股票新闻数据来对股票进行预测方面的研究变得很有意义。

1.2 国内外研究概况

股票市场是十分复杂的系统，其主要的预测方法有传统的机器学习方法，如决策树算法、Boost 算法、SVM 算法等；还有基于深度学习的方法，如卷积神经网络、长短期记忆网络等；以及基于数据挖掘的方法。

1.2.1 基于 SVM 的预测模型

黄秋萍、周霞等人[8]比较了 BP 神经网络、小波神经网络和 SVM 神经网络在股票预测上的表现，发现 SVM 的结果波动性最小，并且三种方法均无法获得稳定的收益。

李坤、谭梦羽[9]提出了优化的支持向量机模型分别预测 13 类股票，发现用小波核替换高斯核能够提高模型的预测准确率，其中使用 Marr 小波核的效果最好。

Mei[10]提出 SVM 与 ARIMA 相结合的方式来预测股价的涨跌方向，先使用 ARIMA 模型对目标进行预测，再将误差作为数据输入到 SVM 模型中，最终的输出作为预测股票涨跌的结果。

Alam[11]提出使用果蝇优化算法来进行参数整定，并提取数据的全局和局部特征作为多核 SVM 模型的输入，来预测股票价格的走势，取得良好效果。

邬春学、赖靖文[12]通过股价走势准确率和股票总盈利作为 SVM 模型的评价指标，探究了 SVM 模型在股票预测方面的可能性，指出 SVM 在股票涨跌预测方面效果一般，还存在一定的改良空间。

黄同愿、陈芳芳[13]针对中国银行股票的价格变化趋势进行预测，比较了不同核函数对于 SVM 模型在股票预测上的影响，最终选定使用径向基核函数。

Huang[14]基于遗传算法改进了支持向量回归机选股模型，对特征提前进行了筛选，提高了模型结果的预测准确率。

李辉、赵玉涵[15]提出逐层提取特征的方法，逐步得到最优特征子集，然后结合 SVM 模型进行预测。该方法可以有效提取与预测相关的特征，提高模型运行效率。

1.2.2 基于决策树的预测模型

Zhou[16]通过采用原始价格数据和 12 个技术指标来提取股票指数所包含的信息，提出将逻辑回归模型与决策树级联起来，并进行模拟交易，实验取得良好的实际收益。

王领、胡扬[17]利用单指标对股票预测有局限性的问题，提出使用 C4.5 决策树算法组合多个指标进行预测，实验结果表明该方法可以提高股票最终收益。

王禹、陈德运等人[18]提出级联多颗 Cart 决策树和 boosting 方法的模型，考虑了数据纵向的关联性，以解决股票预测中的过拟合问题，较好的提升了预测准确率。

1.2.3 基于 Boost 方法的预测模型

陈宇韶、唐振军等人[19]采用皮尔森相关系数分析法从原始数据中提取影响股票涨跌趋势的特征，配合 XGBoost 方法对股票进行预测，在对比实验中该模型具有最低的误差。

王燕、郭元凯[20]探究股票短期预测的方法，提出一种基于 XGBoost 方法的股票预测模型，并利用网格搜索进行参数优化，该模型在众多对照组中有着最高的拟合度。

Zhang[21]引入 Adaboost 集成算法对 2011-2015 年 a 股市场所有公司的年股票收益率进行预测，提高了股票预测的性能并提升了预测准确率。

Zhang[22]提出结合 AdaBoost、概率支持向量机和遗传算法构造了一种新的集成方法来进行股票价格涨跌拐点的预测，该方法在股票投资模拟中获得了较好的收益，显著提高了分类性能。

1.2.4 基于卷积神经网络的预测模型

张贵勇[23]提出 CNN（Convolutional Neural Networks）与 SVM 相结合的股票预测模型，模型先使用卷积神经网络对股票数据进行特征提取，将结果作为 SVM 模型的输入，对股票指数进行和预测，提高预测的精度。

胡悦[24]指出张贵勇所提出的模型最终会陷入局部极小值，输出结果为前些交易日的收盘价，随后提出基于卷积神经网络构建择时模型，对股票涨跌变化点进行预测，实验表面该模型能提升泛化能力。

Hakan[25]提出使用实例和特征之间的相关性被用来对特征进行排序，然后其作为输 CNN 模型的输入，实验表明这种特征选择方法可以减少训练时间和模型复杂度。

陈祥一[26]提出将股票价格涨跌趋势按幅度划分为四个程度，建立卷积神经网络模型，并把一维时间序列特征扩展到二维，使输入数据更能表现出内在的信息，模型具有一定的实际意义。

Catalin[27]设计了两种深度学习模型：长短期记忆网络模型和卷积神经网络模型，实验表明根据这两种模型的预测结果提出交易策略所产生的收益完全不同。

Kim[28]提出基于特征融合的长短期记忆卷积神经网络，模型将提取股票时间序列数据和图片图像数据，实现证明融合的预测模型在股票价格预测上优于单一模型。

1.2.5 基于长短期记忆网络的预测模型

彭燕，刘宇红等人[29]针对 Apple 公司的股票价格使用 LSTM 网络进行预测，以不同的参数进行对比试验，实验结果表明该模型计算法复杂度较小，可以提高预测准确率。

陈佳、刘冬雪等人[30]提出股票数据的特征工程“三步法”来提升 LSTM 模型的预算速度，并且与对照组相比该预测模型的准确度也有明显的提升。

史建楠、邹俊忠等人[31]提出 DMD-LSTM 模型来对股票价格进行预测，用 DMD 算法那提取的股票特征作为 LSTM 的输入，可以更准确的预测出股票的价格涨跌走势。

曾安、聂文俊[32]提出双向 LSTM 的股票预测模型，并结合 Dropout 来提高模型的泛化能力，实验证明该模型可以有效减小误差指标。

王子玥、谢维波、李斌等人[33]将变步长集成方法应用到双向长短期记忆网络中，可训练多个 BLSTM 神经网络，实验证明在变步长条件下，对于特定的股票收益水平得到明显的提升。

Jayanth Balaji 等人[34]探究了多种深度学习模型在股票预测上的效果，实验证明在特定股票上双层 LSTM 模型的效果最好。

1.2.6 基于数据挖掘的预测模型

Ryo Akita [35]提出一种将报纸文章转换成段落向量表示的方法，并用 LSTM 网络对多家公司股票的开盘价格进行建模预测，取得良好的预测结果。

黄丽明、陈维政等人[36]通过多种方法提取新闻特征，并提出多路 RNN 的方法单独处理每个特征，最后通过多层感知器将特征拼接，输出预测结果。实验结果表明该方法可充分提取输入的数据信息，有较好的结果。

朱梦珺、蒋洪迅等人[37]通过匹配知网发布的“情感分析用词集”和特定规则来量化微博文本，并探究了微博情感值和股票价格的你和效果，指出股票上升期的拟合效果大于下降期大于平稳震荡期。

黄润鹏、左文明等人[38]采集微博情绪信息，使用 ROST-CM 算法将文本信息代表的情绪划分 7 种级别，并将挖掘出来的情绪特征作为 SVM 模型的输入之一，对股票价格走势进行预测，结果表明加入该信息可以提高模型的预测准确率。

董理、王中卿等人[39]自行构建构建了股票市场的情感词典，研究了情感分类结果信息、股票技术指标信息、词信息和情感词信息等特征对预测结果起的作用，实验表明情感分类结果和技术指标信息对结果有主导作用。

Misuk Kim 等人针[40]对相关企业 8000 余份财务报告利用词向量的方式进行情感分析，探究这些文本信息代表的情绪与对应股票价格波动的关系，实验证明股价走势的方向随着报告情绪的极性而相应地发生变化。

1.2.7 对研究现状的分析

纵观 1.2.1 至 1.2.6 中前人的研究成果，在一定程度上能够提高对股票预测的效果，但仍存在一定的不足。

（1）股票数据的波动性较强，传统的机器学习算法在对于时间序列的信息提取上有一定缺陷，很难克服股票数据的非线性、混沌性和高噪音的特性。而随着深度模型框架的完善，越来越多研究者使用深度学习的方法来进行股票预测。

（2）使用卷积神经网络的方法需要通过层叠来扩大感受野来捕捉股票，需要耗费更多的资源来提取股票的内在信息，存在算力上的浪费。并且在池化层会丢失大量有价值的信息，忽略局部与整体之间的关联性。

（3）LSTM 是 RNN 的一个优秀的变种模型，非常适合处理与时间序列高度相关的问题。但长距离的信息还是会被弱化，不能对输入的信息的重点进行有效得捕捉，并且不能并行计算。而当下兴起的 Attention 机制可以很好得解决这一问题。

（4）大部分的模型仅仅考虑了单一的股票交易信息的因素，只提取了如股票的收盘价、成交金额、成交量等因素的特征。但股票往往也会受到市场上其他因素的影响，如政府政策、新闻舆论等文本数据所代表的的信息。可以通过提取这方面的文本信息特征来提升模型的预测效果。

1.3 论文的主要研究内容

当今世界正处于信息化的时代，信息的更新迭代非常迅速。企业的股票发生涨跌变化以及当天的交易数据和财经新闻内容所代表的的公司信息都会影响投资者或投资机构对该公司股票的判断，进而引发一系列的买进和抛售行为，导致该公司的股票价格变动；另一方面股票价格变动也会影响公司决策者在公司层面做出相应反映。所以股票的历史数据和与股票相关的财经新闻等信息在一定程度上可以体现出该股票的涨跌趋势，本文设计并实现一种股票预测方法，能够提取获得的股票价格数据和股票新

闻数据的内在信息，对股票未来价格的涨跌趋势做出预测，其结果可以给投资者或投资机构一定的参考。

具体的研究内容主要为以下三个方面：

（1）实现股票相关数据的获取方案。采集包括一段交易日期内的所有股票交易信息和同一段时间内的股票新闻数据。本文使用 Tushare 财经数据包进行股票价格数据采集，使用 Selenium 和 Beautiful Soup 进行新闻数据的采集，并存储到 MongoDB 中。

（2）实现数据的特征化方案。如对股票数据进行归一化，对新闻文本数据提出基于朴素贝叶斯算法的情感量化模型。包括中文分词，TF-IDF 提取特征词等文本处理流程。

（3）实现预测股票涨跌趋势的模型。探索新闻和股票价格特征对股票涨跌趋势的关系。提出一种基于 LSTM-Attention 网络的模型，并通过一些评价指标对其输出结果进行分析。

1.4 论文的组织结构

根据上述的研究内容以及基于新闻情感量化和 LSTM 网络的股票预测模型研究与实现，本文的主要组织结构如下：

第一章绪论，对股票预测模型的北京以及研究意义进行了阐述，并介绍了国内外股票预测模型的研究现状，并对该论文的主要研究内容和主要组织结构进行了介绍。

第二章特征工程及指标构建设计。提出了使用 Tushare 和基于网络爬虫的数据采集方法。并设计了股票数据的归一化方法和基于朴素贝叶斯的新闻情感量化的模型。最后对特征化后的数据进行指标构建来定义模型的输入。

第三章股票预测模型的设计。主要阐述了基于 LSTM-Attention 网络的股票预测模型的构建方法，对每层网络进行了伪代码阐述，以及对模型参数学习进行了定义。

第四章实验。详细讲述了实验的过程，包括数据的采集、股票价格数据的归一化处理和新闻文本数据的情感量化和预测模型的训练，最后以不同模型和不同特征设置了对照组进行对比试验，并对各个实验组的预测结果进行分析。

第五章总结与展望。对本论文进行的工作进行了总结，并说了预测模型存在的不足以及可以改善和进一步研究的方向。

2 股票数据的特征化及指标构建

2.1 数据采集

本文数据主要涉及到所一段交易日内预测股票的股票交易信息和雪球网上该股票的新闻信息。股票的交易数据采集是通过调用 Tushare 财经接口进行获取。股票相关新闻文本的获取主要通过使用 Python 库 Selenium 和 Beautiful Soup 工具进行采集。

作为一个免费的 python 财经数据接口包，Tushare 通过对收集到的包括股票在内的各种金融数据进行清洗、加工并将其存储，从而提供高效、简洁且多样化的易分析数据。

Selenium 是一套完备的 Web 应用测试系统，其可以模拟真实浏览器进行各种操作，主要用以解决 JavaScript 渲染问题，并且能能够获取到浏览器内核已经渲染完毕的 HTML DOM 树及各层 DOM 节点。

Beautiful Soup 是用于网络爬虫中抓取数据的一个库，其主要功能是提供快速处理 HTML 中搜索、分析 DOM 树的方法，能为用户快速解析抓取的网页并获得所需要的数据。

2.1.1 采集股票数据

	ts_code	trade_date	open	high	low	close	pre_close	change	pct_chg	vol	amount
0	000001.SZ	20191231	16.57	16.63	16.31	16.45	16.57	-0.12	-0.7242	704442.25	1154704.348
1	000001.SZ	20191230	16.46	16.63	16.10	16.57	16.63	-0.06	-0.3608	976970.31	1603152.786
2	000001.SZ	20191227	16.53	16.93	16.43	16.63	16.47	0.16	0.9715	1042574.72	1741473.179
3	000001.SZ	20191226	16.34	16.48	16.32	16.47	16.30	0.17	1.0429	372033.86	610381.757
4	000001.SZ	20191225	16.45	16.56	16.24	16.30	16.40	-0.10	-0.6098	414917.98	679664.596
5	000001.SZ	20191224	16.23	16.50	16.23	16.40	16.24	0.16	0.9852	459128.42	752351.618
6	000001.SZ	20191223	16.68	16.68	16.17	16.24	16.59	-0.35	-2.1097	715792.72	1173372.080
7	000001.SZ	20191220	16.55	16.68	16.44	16.59	16.55	0.04	0.2417	644478.38	1067869.779
8	000001.SZ	20191219	16.55	16.74	16.44	16.55	16.46	0.09	0.5468	675536.76	1119925.969
9	000001.SZ	20191218	16.43	16.66	16.41	16.46	16.50	-0.04	-0.2424	797091.33	1317608.543
10	000001.SZ	20191217	16.00	16.63	15.98	16.50	16.13	0.37	2.2939	1203104.90	1969891.323

图 2-1 平安银行股票数据

使用 Tushare 接口获得所需要的数据非常便于操作进行数据分析和可视化，因为其返回值都是 DataFrame 的形式。本文将通过条用 Tushare 接口来获取股票数据存储在本地，之后再进行后续处理工作。以平安银行（000001.SZ）为例，获取的数据及其字段如图 2-1 所示。其中各字段的含义如表 2-1 所示。

表 2-1 tushare 返回字段含义表

字段名	含义	字段名	含义
ts_code	股票代码	pre_close	前一交易日收盘价
trade_date	交易日期	change	涨跌
open	开盘价	pct_change	涨跌幅
high	当日最高价	vol	成交次数
low	当日最低价	amount	成交金额
close	收盘价		

2.1.2 股票数据存储结构

通过 Tushare 获取的股票交易数据是 DataFrame 形式的，可非常方便的通过 Python 的 Pandas 和 Numpy 工具包来进行数据操作，但需要经过序列化后，才能存储到数据库中。MongoDB 可以很方便的通过 JSON 来进行数据的序列化，通过一系列的 Key-Value 键值对来表示数据，符合大众的阅读习惯，并且 MongoDB 是专门为可拓展性、高性能和高可用而设计的数据库，非常适合敏捷开发。因此本文采用 MongoDB 作为存储数据的数据库。

每个企业的股票交易数据都分别用 MongoDB 的一个 Collection 进行存储，表名统一以“股票代码_Price”（如 000001SZ_Price）的格式进行定义。该表存储本股票的每日行情数据。表的具体定义如表 2-2 所示， date 字段表示交易日期，使用唯一索引约束；open 字段表示当日开盘价；high 表示当日最高价；low 表示当日最低价；close 表示当日收盘价；change 表示当日涨跌；pct 表示当日涨跌幅；vol 表示当日成交量；amount 表示当日表示当日成交金额。

表 2-2 股票数据表

字段	类型	可否为空	索引	注释
date	Date	否	唯一索引	股票代码
open	Date	是		当日开盘价
high	Double	是		当日最高价
low	Double	是		当日最低价
close	Double	是		当日收盘价
change	Double	是		当日涨跌
pct	Double	是		当日涨跌幅
vol	Double	是		当日成交量
amount	Double	是		当日成交金额

2.1.3 采集新闻数据

雪球网（xueqiu.com）因为其独特的产品设计和运营策略，并能够为投资者提供股票数据查询、新闻订阅和交流信息等服务，雪球 App 成为国内高质量投资者的聚集地，并在股票应用市场拥有强大的影响力。本文使用 Selenium 和 Beautiful Soup 来爬取雪球网中新闻数据。

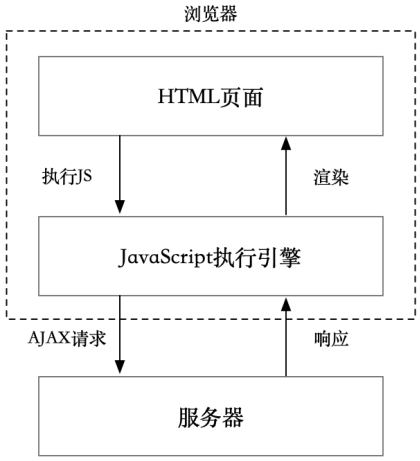


图 2-2 AJAX 原理

随着 Web 技术的发展，越来越多的网站采用 AJAX（Asynchronous JavaScript And XML）技术来构建网站页面。AJAX 是一种异步的 JavaScript 与 XML 技术，其原理是客户端请求得到完整的 HTML 页面并在 JavaScript 执行引擎中运行页面中的脚本代码，然后仅向服务器发送请求必要的数据，如此可以减少客户端与服务器之间交换的数据量。然后等待异步回调得到所请求的数据，收到 Response 包后将取得的数据填充到 HTML 页面中。这就是页面渲染的过程，原理如图 2-2 所示。而雪球网就是应用了 AJAX 这项技术。

由于 Selenium 能够完成网页的渲染过程，并执行 JavaScript 代码，所以在数据采集中使用该框架，可以无需人工分析 AJAX 的请求逻辑，自动渲染使用 AJAX 技术构建的网站，获取到完整的 HTML 页面数据。Beautiful Soup 有许多内置方法，可以很方便的解析 DOM 树，根据用户制定的规则提取出所需要的数据。

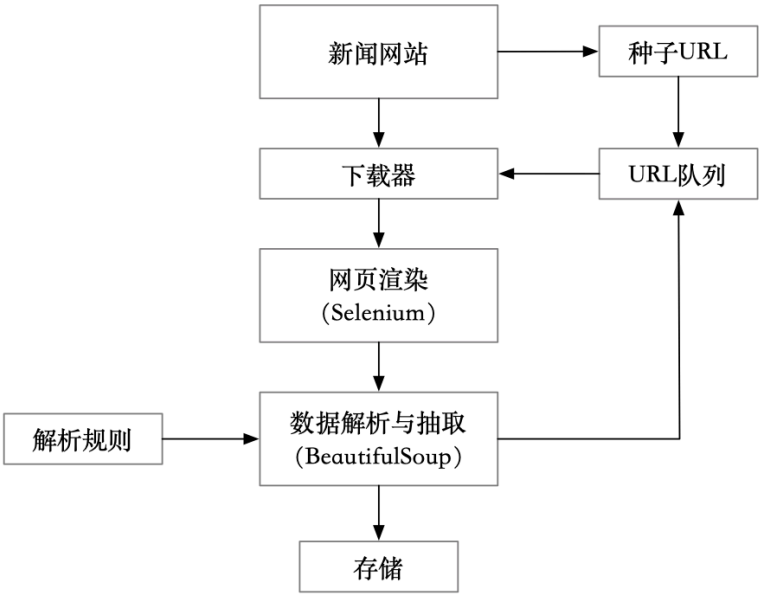


图 2-3 爬虫方案流程图

股票新闻的数据采集流程如图 2-3 所示。其步骤如下：

（1）从雪球网的特定股票新闻专区中提取种子 URL，将其放入队列中；

（2）下载器每次从 URL 队列中依次取出 URL，向目标服务器发出 Request；

（3）获得到服务器返回的 Response 包后，Selenium 模拟浏览器渲染 HTML 页面。期间会执行嵌入 HTML 页面中的 JavaScript 代码，与目标服务器多次发起请求数据过程，得到所需数据并嵌入 HTML 代码中，最终渲染出来；

（4）通过分析雪球网新闻专区的网页结构得到的解析规则，使用 Beautiful Soup 解析 HTML 抽取对应的新闻网站的内容；

（5）将采集的新闻数据存储到数据库中。

2.1.4 新闻数据存储结构

新闻数据的存储也同样使用 MongoDB 来实现。与股票交易数据的存储结构类似，根据不同股票代码，建立相应的 Collections， Collection 中的每条记录的就代表着特定股票当天的所有新闻数据。date 表示新闻发表的日期，使用唯一索引约束；news 代表每天的新闻数组，囊括了当天所有的新闻数据，新闻数据表的存储字段如表 2-3 所示。News 字段的结构以及类型如表 2-4 所示，包括新闻标题、新闻正文、新闻原始链接等。

表 2-3 新闻数据表

字段	类型	可否为空	索引	注释
date	Date	否	唯一索引	新闻日期
news	Array	是		当日新闻数据

表 2-4 news 字段存储结构及类型

字段	类型	可否为空	注释
title	String	否	新闻标题
content	String	是	新闻正文
URL	String	是	原始链接

2.2 股票数据处理

现实中收集到的数据由于各种各样的原因，可能会出现有缺失值、类型不一致值甚至是溢出值等问题。这些异常的数据如果不提前进行处理，直接使用到模型当中会

严重影响执行效率，甚至会出现难以察觉的错误。所以数据的预处理工作就显得尤为重要。而股票价格数据的预处理主要包括数据的缺失值处理以及数据的归一化处理。

2.2.1 数据的缺失值处理

在通过调用 Tushare 获取的股票数据和爬虫获取的新闻数据中，可能由于某些原因，例如网络延迟或者编码错误，有些信息发生丢失。如果数据丢失了大量有用信息，收集到的数据集就不能代表股票市场真实的信息，对预测结果产生不良的影响。处理缺失值常用的方法有如下几种。

（1）删除记录。该方法是直接将有缺失值的数据直接删除，一般来说除非是无法填补缺失的数据，不会轻易删除记录，因为该方法会改动原本记录中包含的其他的有用信息。

（2）手工填补。当缺失的记录很少时，可以采用该方法人工填补缺失值的真实值。

（3）采用默认值、均值、中位数等填补。对特定类型数据可以设置默认常数来填补，或者用该数据字段的总体平均数或其他均值类型来填补。该方法可以保留其他非缺失值的有效信息。

（4）使用最可能的值填充缺失值。可以通过一些基本的数据挖掘方法，如回归分析、贝叶斯方法或决策树来预测该缺失值的真实数据。该方法会比方法（3）更贴近真实值，但如果缺失值过多，就不适用该方法。

对于股票的每日交易数据，因其为时间序列数据，某天的数据会跟相邻几天的数据相似，所以采用最近邻插补法。取前一日期的有效数据和后一有效日期的均值来进行填补。对于股票的新闻数据，因为其为文本信息，若有缺失值不能采取其他方式来填补，所以采用删除记录的方法。

2.2.2 数据的归一化处理

数据归一化处理是数据挖掘过程中普遍要执行的的一项工作，也是数据预处理中的一个重要环节。因为不同字段的数据往往具有不同的量纲和单位，若数值之间相差

太大，会影响模型的收敛速度以及预测的结果。因此在模型计算之前，需要对数据进行归一化操作。归一化方式主要有以下几种:

（1）Max-Min 归一化

把每一个数据与该特征下最小值的差值去除以最大值和最小值的差值得到的结果作为归一化后的数值，该特征下的数据由此被映射到[0,1]之间，其转换公式如式 2-1 所示。

$$normal = \frac{x - min}{max - min} \quad \text{（式 2-1）}$$

（2）Z-Score 归一化：

Z-Score 归一化也称作标准差归一化，其过程与正态分布的标准化过程一致，其转化公式如式 2-3 所示，其中，x 代表原始的数据，μ代表原始数据的均值，σ表示原始数据的标准差。经过 Z-Score 归一化处理后，数据的均值变为 0 标准差变为 1。

$$normal = \frac{\mathrm{x} - \mu}{\sigma} \quad \text{（式 2-2）}$$

Max-Min 归一化，在将数据缩放至 0 到 1 的过程中，参考的是数据中的最大值和最小值作为限定标准，在数据采集过程中可能会出现数据溢出的情况，导致这两个值变得非常极端，进而影响整体的数据，导致模型预测结果不准确。为了避免这种情况的发生，本文选用 Z-score 归一化方法。

2.3 新闻的情感量化

情感量化是对带有感情色彩和主观性倾向的文本数据进行分析的过程，即分析文本信息所代表的情感倾向是正面还是反面，对某个观点是支持还是反对。它与传统的文本主题分类又不相同，传统主题分类是分析文本讨论的客观内容，而情感分类是要从文本中得到它是否支持某种观点的信息。比如，“专家：经历这次疫情纳斯达克指数还能大涨，这种说法颇具讽刺的以为。”传统主题分类会将这句话归类为“财经主题”，而情感分类则要挖掘出专家对于“经历这次疫情纳斯达克指数还能大涨”这个

观点持反面态度。这是一项具有较大使用价值的分类技术，可以将股票相关消息的情感元素作为预测股票涨跌的依据。

一般而言，新闻对于股票价格的影响可分为利好和利空两类。利好新闻是指有利于股价上涨的消息。比如上市公司收购其他公司、公司项目中标、财报中净利润大幅增长等消息，这些消息反映出上市公司盈利能力改善，人们对公司股价的预期也会上升。相反，利空新闻是指会影响股价下跌的消息。比如上市公司经营业绩不佳、内幕丑闻、与其他企业财务纠纷等消息，这些消息会导致人们抛售公司的股票。由此可见，股票预测中新闻利好与利空的判断，可以用分类算法来进行分类。本文使用朴素贝叶斯方法来进行新闻的情感分类。基本流程如图 2-4 所示。

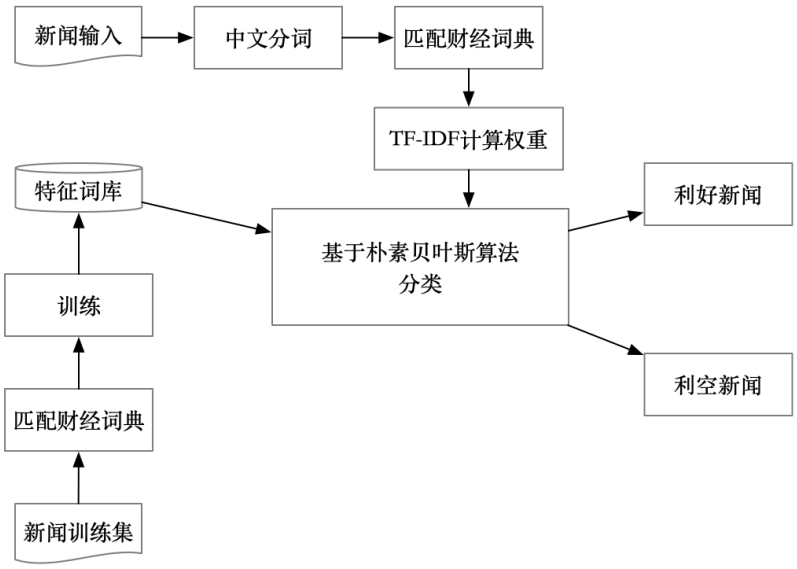


图 2-4 基于朴素贝叶斯的情感量化流程

2.3.1 中文分词

文本分词方法要根据语言类型来制定，常见的有中文和英文两种。对英文文本进行分词的方法比较简单，主要是用非字母符号将单词隔开，如标点符号、空格等。然而对于中文分词，这个过程是比较复杂的。因为中文词并不像英文词那样有明显界限，也没有一个确定的分词规范。并且中华文化博大精深，很多词有丰富的语义，歧义词

的存在非常普遍，甚至人工来切分也需要领悟下文的语义。所以中文分词方法也多种多样，同一句话用不同的方法会产生不同的效果，而分词精度是影响模型准确率的重要因素。

现在的开源中文分词工具或者模块已经很丰富，并且很多都有一些在封闭测试集上的效果对比数据，不过这仅仅只能展现这些分词工具在这个封闭测试集上的效果，并不能全面说明问题，所以选择一个适合自己业务的分词器可能更重要。经过比较本文选用 pkuseg 来进行新闻的中国分词。

pkuseg 是由北京大学语言计算与机器学习研究组开发的中文分词工具，与其他的分词相比有如下几个特点：

（1）多领域分词。不同通用的中文分词工具，此工具包可以让用户根据待分词文本的领域特点选择不同的模型，目前覆盖了新闻、网络领域、医药领域和旅游等领域，并且支持混合领域的分词预训练模型。

（2）更高的分词准确率。相比于其他的分词工具包，当使用相同的训练数据和测试数据，pkuseg 可以取得更高的分词准确率。

（3）支持用户自行训练模型。支持用户使用自己的的标注数据来进行训练，得到特定领域的分词模型。并且支持词性标注。

2.3.2 匹配财经词典

THUOCL（THU Open Chinese Lexicon）是由清华大学自然语言处理与社会人文计算实验室整理的一套高质量中文词库，其中 THUOCL_caijing.txt 包含了 3830 条财经词汇，词频统计语料库来源于新浪新闻。

2.3.3 TF-IDF 提取特征词

TF-IDF 是一种信息检索的统计方法，其核心思想为：如若某个词或短语在一篇文章中有很高的出现的频率 TF，并且这个词在其他文章中出现次数较少，则认为该词或短语的类别区分能力较强，适合用来分类。

TF（Term Frequency）表示词频，表示某个词在文章中的出现次数，因为文章长短各不相同，为了保证不同长度的文章之间可以进行比较，会对词“词频”进行归一化处理，如式 2-3 所示。

$$\mathrm{TF} = \frac{\text{某个词在文章出现的次数}}{\text{文章的总次数}} \qquad \text{（式 2-3）}$$

IDF（Inverse Document Frequency）表示逆向文件频率，如果文档中出现某词条的数量越少，则说明该词条区分类别的能力较强，IDF 就越大。IDF 的数学表示如式 2-4 所示，因为分母不能为零，避免类别里所有文档都不包括该词的情况，分母应定义为包含该词的文档数加 1；分子定义为语料库的文档总数。将分式的结果取对数是为了避免结果太小。

$$\mathrm{IDF} = \log\left(\frac{\textit{语料库的文档总数}}{\textit{包含该词的文档数} + 1}\right) \qquad \text{（式 2-4）}$$

TF-IDF 表示词频-逆文件频率，定义如公式 2-5。根据公式可以看出，TF-IDF 与某词在文章中出现的次数成正比，与包含该词的文档数成反比。

$$\text{TF-IDF} = \text{TF} * \text{IDF} \qquad \text{（式 2-5）}$$

所以，自动提取特征词的步骤就是计算出文档的每个词的 TF-IDF 值，然后根据计算的 TF-IDF 值进行降序排列，取排在最前面的几个词。

2.3.4 朴素贝叶斯算法

朴素贝叶斯分类是一种非常简单有效基于概率的分类算法，被广泛得应用于各个领域。其核心思想是在贝叶斯算法的基础上，做出“样本数据集的属性和类别之间不存在依赖关系”的假设，可以简化计算。

（1）贝叶斯基本原理

贝叶斯算法是根据已经发生的事件来预测未来某事件发生概率的一种算法。如果知道已经发生的事件的概率，那么可以根据推导出未发生的事件出现的概率。为更好的用数学语言来描述该原理，需要先定义几个公式。

条件概率公式如式 2-6 所示。其中等式左边的$P(B|A)$代表事件 A 已经发生后事件 B 发生的概率；等式右边的分母 P(A)代表事件 A 发生的概率，分子 P(AB)代表事件 A、B 都会发生的概率。

$$P(B|A)=\frac{P(AB)}{P(A)} \quad \text{（式 2-6）}$$

将式 2-6 做简单的等式变换可得到联合概率公式，如式 2-7 所示。

$$P(AB)=P(A)P(B|A)=P(B)P(A|B) \quad \text{（式 2-7）}$$

全概率公式如式 2-8 所示。设实验的样本空间为 S，A 为随机实验的一个其中一个结果，事件 $B_1, B_2, \dots, B_n$ 是 S 的一个划分，且$\sum_{i=1}^{n} \mathrm{P}(B_i)=1$。

$$\mathrm{P}(\mathrm{A})=\mathrm{P}(\mathrm{A}|B_1)\mathrm{P}(B_1)+\mathrm{P}(\mathrm{A}|B_2)\mathrm{P}(B_2)+\cdots+\mathrm{P}(\mathrm{A}|B_n)\mathrm{P}(B_n) \quad \text{（式 2-8）}$$

假设测试 E 的样本空间为 S，X 为 E 的一个事件，事件 $B_1, B_2, \dots, B_n$ 是 S 的一个划分，且$\mathrm{P}(A)>0$，$\mathrm{P}(B_i)>0$，则有贝叶斯公式如式 2-9 所示。

$$P(B_i|X)=\frac{P(X|B_i)P(B_i)}{P(X)}=\frac{P(X|B_i)P(B_i)}{\sum_{i=1}^{n}\mathrm{P}(X|B_i)\mathrm{P}(B_i)} \quad \text{（式 2-9）}$$

通俗的来说，贝叶斯公式就是在已知所有$P(B_i)$（即事件B_i发生的先验概率）和所有$\mathrm{P}(X|B_i)$（即事件B_i已经发生后事件X发生的条件概率），来求未知的$P(B_i|X)$（即未来事件 X 发生时，事件B_i也会发生的后验概率）的方法。在一般的分类模型中，B_i代表着时间的类别，而在股票涨跌预测中就只有涨和跌两个类别。

（2）朴素贝叶斯原理

对于一个新闻事件 X，在经过中文分词之后该时间可用特征向量来表示，即$\mathrm{X}=(x_i, x_2, \dots, x_n)$。根据贝叶斯公式中的$P(X|B_i)=P(x_i, x_2, \dots, x_n|B_i)$很难求出，这是一个超级复杂的有 n 个维度的条件分布。朴素贝叶斯原理就是在贝叶斯原理的基础上对给定事件的属性值之间做出相互独立的假设，即$x_i, x_2, \dots, x_n$之间相互独立。根据条件独立公式，贝叶斯原理可被等价于式 2-10 的形式。其中$P(x_j|B_i)$可以由中文分词后的数据统计得到，即类别B_i中出现中文词x_j的概率。

$$P(X|B_i) = P(x_i, x_2, \dots, x_n|B_i) = \prod_{j=1}^{n} P(x_j|B_i) \quad \text{（式 2-10）}$$

实际应用场景中各属性之间北非完全的独立关系，所以该假设在一定程度上会降低了贝叶斯分类算法的分类效果，但是该假设能极大地简化贝叶斯方法的复杂性，对于组成新闻文本的各中文词语来说，其相关性较小，所以可以用该方法在不过多影响预测准确率的情况下减小模型计算量[41]。朴素贝叶斯公式如式 2-11 所示。

$$P(B_i|X) = \frac{\prod_{j=1}^{n} P(x_j|B_i)\,\mathrm{P}(B_i)}{\sum_{i=1}^{n} \prod_{j=1}^{n} P(x_j|B_i)\,\mathrm{P}(B_i)} \quad \text{（式 2-11）}$$

（3）朴素贝叶斯量化器

对于我们预测的类别结果 C 是使$P(B_i|X)$数值最大所代表的的那个类别，即将所有类别的后验概率计算出来，取最大值。那个最大值所代表的的类别就是预测的结果。

最终新闻这类文本类信息会被量化成数值 c（值为-1 或 1），其中 c 为情感指标 1 代表新闻是利好，-1 代表利空。

2.4 指标构建

2.4.1 股票价格特征构造

由 Tushare 获得的数据有很多字段，股票价格特征是股票预测问题中最直接的、最直观的特征；相对于股价的绝对数值，股价的涨跌幅对于股价涨跌预测更为有效；而股票的成交量和成交金额在某种程度上能显示该股票的热门程度。

不同的股票，由于股票价格基数不同，其股价的绝对值差异往往很大，所以对收盘价、涨跌幅、成交量和成交金额等特征进行 Z-Score 归一化，将这些归一化之后的特征序列作为股票价格方面的特征。形式化描述股票在第 i 个交易日价格方面的特征如式 2-12 所示：

$$q_i = \text{<close, ptc\_chg, vol, amount>} \quad \text{（式 2-12）}$$

式子中定义的股价涨跌幅序列是股票价格方面的特征，作为后续 LSTM 预测模型的输入。

2.4.2 新闻情感特征构造

在 3.3 节新闻的情感量化中，通过使用 THUOCL 匹配新闻领域进行中文分词，然后进行 TF-IDF 的特征向量化，对情感强烈的词进行人工提高权重。最后通过朴素贝叶斯算法进行概率预测得到新闻的类别以及概率。故实验中，本文最终将新闻这类文本类信息会被量化成 c 值，其中 c 为情感指标 1 代表新闻是利好，-1 代表利空。

形式化，某天某只股票的 n 个新闻文本被量化成具体的数值 s，如式 2-13 所示。

$$s = \frac{\sum c_i}{n} \qquad \text{（式 2-13）}$$

式子中定义的新闻情感值 s 最终进行 Z-Score 归一化之后是新闻情感方面的特征，作为后续 LSTM 预测模型的输入。

2.4.3 模型指标

本文主要研究股票涨跌预测问题。股票在每个交易日结束后，都会留下一些关键交易数据：收盘价、涨跌幅、成交量和成交金额等。用这些特征来代表这一交易日的股票价格。在持续的一段交易日内，股票会产生一个收盘价格的序列，记这段持续的价格序列为:

$$q_1, q_2, \dots, q_T$$

其中p_t表示股票在第 t 个交易日的收盘价格，T 为这段时间的交易日数量。

除了股票的价格信息，股票相关的市场信息也是股票预测的重要依据。假设在这段交易日内，该只股票相关的市场信息(相关新闻、消息等)为一组文本序列，每个交易日对应一个文本,经过情感量化之后，本文记这段持续的新闻情感序列为：

$$s_1, s_2, \dots, s_T$$

股票价格相对前一个交易日会有所涨跌，本文采用收盘价格作为涨跌的判断依据，定义股票在第 t 个交易日的涨跌情况为yt，若其值为 1 代表价格上涨，若其值为 0 代表下跌或持平。于是股票涨跌预测问题定义如下:

已知一段持续的交易日内某只股票的价格序列和相关新闻序列:

$$q_1, q_2, \dots, q_T \quad s_1, s_2, \dots, s_T$$

预测该只股票下一个交易日的收盘价的涨跌情况，即预测y_{T+1}。可以看出，股票预测问题是利用股票历史信息预测未来涨跌的问题。

2.5 本章小结

本章详细介绍了研究所需要的数据的采集、数据的特征化以及指标的构建。2.1 节介绍了数据采集的方法，使用 Tushare 接口和 Selenium、Beautiful Soup 爬虫工具来分别获取股票价格数据和股票新闻数据。2.2 节介绍了股票数据的处理方法，使用 Z-Score 方法来做数据的归一化。2.3 节给出了基于朴素贝叶斯算法的新闻情感量化方法。最后 2.4 节构建了模型所需要的特征以及模型指标。

3 股票预测模型的设计

在股票涨跌预测问题中，不论是股票每日的价格信息，还是相关的市场信息，都具有时间维度上的顺序。股票每个交易日会产生收盘价，股票的相关新闻也能对应到某个交易日。相对于其他的分类算法，如支持向量机、逻辑斯蒂回归，循环神经网络模型能够自然地融入时间这一维度，描述特征在时序上的关联。

除了时序性，股票预测问题还有延迟性与持续性特征。首先从相关新闻的发布到新闻影响股价的波动有一定的时间间隔。在新闻发布后，人们从发现新闻，传播新闻，到分析新闻，做出决策需要时间，新闻信息作用到股价上需要一段时间。其次新闻对股价的影响往往不局限于单个交易日，而是在一段交易日内产生持续的作用。股票预测延迟性与持续性的特征，要求预测模型能够将新信息保存一段时间，具有一定的“记忆”能力。将入注意力机制的长短时记忆网络如同它的名称一样，能够将输入信息记忆在模型中，是处理股票预测问题的合适模型。

3.1 LSTM 层

普通循环神经网络受到信号随时序指数衰减的影响，无法处理长时间间隔的信息关联，并且容易出现梯度消失等问题。为了解决这一难题，LSTM 网络的每层都有多个相同结构的记忆块组成，每个记忆块，都包含一个“输入门”(InputGate)，一个“输出门”(OutputGate)和一个状态节点。输入门控制输入信息被记忆块接收的比例，输出门控制记忆块输出信息向外界传递的比例，而状态节点则蕴含记忆块所接收的所有历史信息，形成了对输入信息的“记忆”。凭借记忆块的结构，LSTM 网络能够记忆间隔超过 1000 个时间单位的信息。

记忆块（Memory Block）是长短时记忆网络的基本构成单元。记忆块如同大脑中的神经元，具有记忆信息与连接周围神经元的作用。记忆块与时序相关，在 t 时刻它的输入是x_t，输出是 h_t，记忆块“记忆”的历史信息为C_t。每个记忆块都有三个门，即输入门i_t，遗忘门f_t以及输出门o_t。输入门控制 t 时刻输入信息的比例，遗忘门控制 t

时刻记忆历史信息的比例，输出门则决定了记忆块向下一层输出信息的比例。记忆块具有时序的状态，C_t包含了从时刻 0 到时刻 t 的所有输入信息，是记忆块在时刻 t 的状态。 记忆块的结构如图 3-1 所示。

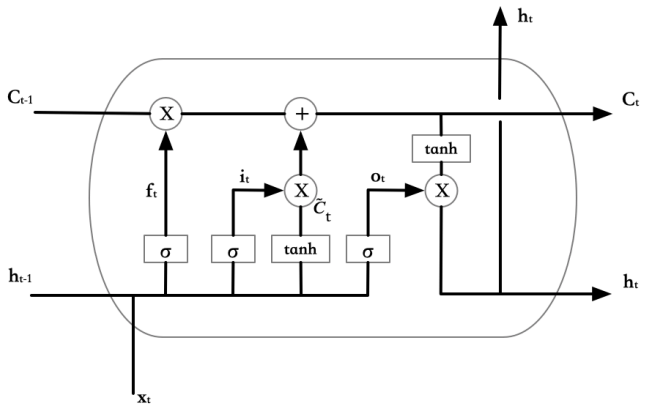


图 3-1 记忆块结构

遗忘门f_t由输入信息x_t和记忆块上一时刻的输出信息 h_{t-1} 决定，经过 sigmoid 函数变换，遗忘门的数值在 0 到 1 之间，如式 3-1 所示。f_t表示记忆块在前一时刻的状态C_{t-1}有多大比例保留到当前 t 时刻。f_t值为 0 表示完全遗忘前一时刻的状态，f_t值为 1 表示完全记忆前一时刻的状态。

输入信息x_t与前一时刻的输出 h_{t-1}，经过 tanh 函数变换，形成当前时刻的状态增量$\tilde{C}_t$，如式 3-2 所示。状态增量$\tilde{C}_t$的值在区间(-1,1)，表示输入信息x_t能够给记忆块的状态造成的增量大小。

与遗忘门类似，输入门i_t的值由输入信息x_t和前一时刻的输出信息 h_{t-1} 决定，如式 3-3 所示。输入门控制着状态增量$\tilde{C}_t$被记忆块接收的比例。如果输入门i_t值为 0，那么状态增量$\tilde{C}_t$会被完全忽略;如果输入门i_t 值为 1，那么$\tilde{C}_t$会被完全计入状态Ct。

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad \text{（式 3-1）}$$

$$\tilde{C}_t = tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \quad \text{（式 3-2）}$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad \text{（式 3-3）}$$

记忆块的更新状态如下。状态值C_t 由两部分组成，一个部分是前一刻的状态C_{t-1}，这个部分由遗忘门控制保留的比例；另一个部分是状态增量$\tilde{C}_t$，这一个部分由输入门决

定接受增量的比例。状态值$\tilde{C}_t$随着时间的更新公式如式 3-4 所示。值得注意的是，公式中的符号“*”表示的是向量之间的按位乘法，即在记忆块中门是对信息的一种缩放。

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \quad （式 3-4）$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad （式 3-5）$$

$$h_t = o_t * \tanh(C_t) \quad （式 3-6）$$

输出门o_t与遗忘门和输入门类似，其值由输入信息x_t与前一刻的输入信息 h_{t-1} 综合而来，如式 3-5 所示。输出门o_t的大小决定了记忆块的状态C_t有多大概率作为输出，被其他的神经网络层捕获。

符号 h_t 表示记忆块在时刻 t 的输出。这一输出基于记忆块的状态C_t，但是需要经过输出门的过滤。状态C_t首先经过 tanh 函数变换，将状态压缩到区间（-1, 1），再由输出门o_t决定输出的比例，如式 3-6 所示。

表 3-1 LSTM 层伪代码

LSTM 层：lstm_layer($\vec{x}_1, \vec{x}_2, \dots, \vec{x}_r$)

输入：一组按时序排列的向量组$\vec{x}_1, \vec{x}_2, \dots, \vec{x}_r$

参数张量$\vec{\varphi} = [W_i, W_f, W_o, W_C, \vec{b}_i, \vec{b}_f, \vec{b}_o, \vec{b}_C]$

输出：按时间排序的向量组$\vec{h}_1, \vec{h}_2, \dots, \vec{h}_T$

过程：

1. $\vec{C}_0 := \vec{0}$
2. $\vec{h}_0 := \vec{0}$
3. **for** t **from** 1 to T:
4. $\vec{i}_t := \sigma(W_i \cdot [\vec{h}_{t-1}, \vec{x}_t] + \vec{b}_i)$ //输入门
5. $\vec{f}_t := \sigma(W_f \cdot [\vec{h}_{t-1}, \vec{x}_t] + \vec{b}_f)$ //遗忘门
6. $\vec{o}_t := \sigma(W_o \cdot [\vec{h}_{t-1}, \vec{x}_t] + \vec{b}_o)$ //输出门
7. $\tilde{C}_t := tanh(W_C \cdot [\vec{h}_{t-1}, \vec{x}_t] + \vec{b}_C)$ //当前输入的单元状态
8. $\vec{C}_t = \vec{f}_t * \vec{C}_{t-1} + \vec{i}_t * \tilde{C}_t$ //当前时刻的单元状态
9. $\vec{h}_t = \vec{o}_t * \tanh(\vec{C}_t)$ //输出值
10. **endfor**
11. **return** $\vec{h}_1, \vec{h}_2, \dots, \vec{h}_T$

LSTM 记忆块是一种时序上的函数变换，它将输入时序序列变换为另一个时序序列。因为记忆块内置状态单元，并且具有输入门与遗忘门的特殊结构，所以记忆块具备关联长间隔信息的能力。

注意到，记忆块的输入x_t和状态C_t都是向量，也就是说，记忆块可以单独作为神经网络的一层。LSTM 网络由一层或多层构成，本文将一层 LSTM 网络的功能用伪代码表示出来，如表 3-1 所示。可以看出，LSTM 层的输入与输出都是时序上的向量组，并且输入向量与输出向量在时间上一一对应。每个时刻，LSTM 层都进行一组相同的操作，计算出记忆块的状态和输出信息。多层 LSTM 模型由多个 LSTM 层拼接而成。

3.2 Self-Attention 层

人们在看东西的时候一般不会从头看到尾全部都看，往往只会根据需求观察注意特定的一部分。受此启发，有关研究者提出了 Attention 注意力机制。简单来说 Attention 注意力机制就是一种权重参数的分配机制，通过关注与输入元素相似的部分协助模型捕捉重要信息。其最大的优势是能一步到位的考虑到全局联系和局部联系，且每步的结果不依赖与上一步，能并行化计算，这在大数据的环境下尤为重要。

Attention 函数的本质可以被描述为一个查询（query）到一系列键值对（Key-Value）的映射。在计算 Attention 时主要分为三步：

（1）相似度计算。计算 query 和每个 key 的相似度或相关性，最终得到权重。其具体的数学表达式 F(Q, K)如式 3-7 所示，分别代表向量点积（Dot Product）、权重（General）、拼接权重（Concat）和感知机（Perceptron）等常用的相似度计算方法。

$$\mathrm{F}(\mathrm{Q}, K_i) = \begin{cases} Q^T K_i \\ Q^T W_a K_i \\ w_a[Q;\ K_i] \\ v_a^T \tanh\left(W_a Q + U_a K_i\right) \end{cases} \qquad \text{（式 3-7）}$$

（2）进行归一化操作，可以通过一个 softmax 函数的特性更加突出重要元素的权重，得到权重值 a_i，其具体的数学表达式如式 3-8 所示；

（3）将权重和对应的 value 进行加权求和得到最后的 attention 值，其具体的数据表达式如式 3-9 所示。

$$a_i = \text{Softmax}(\text{F}(\text{Q}, K_i)) = \frac{\exp\left(\text{F}(\text{Q}, K_i)\right)}{\sum_j \exp\left(F(Q, K_j)\right)} \quad \text{（式 3-8）}$$

$$Attention(Q, K, V) = \sum_j a_i v_i \quad \text{（式 3-9）}$$

Self-attention 是一般 attention 的一种特殊情况，self-attention 是将序列中的每个元素和该序列中所有元素进行 attention 计算。研究者经常把 self-attention 视为一个神经网络层，并且和其他神经网络等配合使用，然后应用于各种任务当中。相比于传统的 RNN 和 CNN，attention 机制具有如下优点：

（1）self-attention 对长距离的依赖关系有一定的捕捉能力，能够学习一个句子的内部结构，既捕捉了全局联系，也关注了元素的局部联系；attention 函数在计算 attention 的 value 时，是进行序列中每一个元素和其他元素的对比，在这个过程中每一个元素间的距离都是 1。相比之下 RNN 是通过一步步递推才能捕捉到全局的联系，并且其捕捉长距离的依赖关系能力较差；而 CNN 则需要通过层叠来扩大感受野，需要耗费更多的资源，且效果一般。

（2）因为 self-attention 机制直接把序列两两比较，时间复杂度是 $O(n^2)$，并且每一步的计算都不依赖上一步的计算结果，可以进行并行计算，从而可以减少模型训练的时间。

通过 3.1 节中 LSTM 层的计算结果可以得到$\vec{h}_{i1}, \vec{h}_{i2}, \ldots, \vec{h}_{it}$的值，即第 i 层 t 时刻的输出值，将这些输出值视为 Self-attention 自注意力层的输入。首先通过点乘计算其第 i 层 T 时刻输出值$\vec{h}_{iT}$与第 i 层所有输出值构成的输出向量H_i间的相似度 $\text{F}(H_i, h_{iT})$，然后通过 Softmax 归一化得到 t 时刻的权重 a_t，其具体数学表达式如式 3-10 所示。

$$a_t = \text{Softmax}(\text{F}(H_i, h_{it})) = \frac{\exp\left(\text{F}(H_i, h_{it})\right)}{\sum_j \exp\left(F(H_i, h_{ij})\right)} \quad \text{（式 3-10）}$$

得到$a_1, a_2, \dots, a_T$权重向量，本文将一层 LSTM-Attention 网络的功能用伪代码表示出来，如表 3-2 所示。

表 3-2 Self-Attention 层伪代码

Self-Attention 层：lstm-attention_layer $(\vec{h}_1, \vec{h}_2, \dots, \vec{h}_i)$
输入：一组按时序排列的向量组$\vec{h}_1, \vec{h}_2, \dots, \vec{h}_i$
参数张量$\vec{\varphi} = [W_i, W_f, W_o, W_C, \vec{b}_i, \vec{b}_f, \vec{b}_o, \vec{b}_C]$
输出：按时间排序的向量组$\vec{A}_1, \vec{A}_2, \dots, \vec{A}_T$
过程：
for t **from** 1 to T：
$\vec{\imath}_t \coloneqq \sigma(W_i \cdot [\vec{h}_{t-1}, \vec{x}_t] + \vec{b}_i)$ //输入门
$\vec{f}_t \coloneqq \sigma(W_f \cdot [\vec{h}_{t-1}, \vec{x}_t] + \vec{b}_f)$ //遗忘门
$\vec{o}_t \coloneqq \sigma(W_o \cdot [\vec{h}_{t-1}, \vec{x}_t] + \vec{b}_o)$ //输出门
$\tilde{C}_t \coloneqq tanh(W_C \cdot [\vec{h}_{t-1}, \vec{x}_t] + \vec{b}_C)$ //当前输入的单元状态
$\vec{C}_t = \vec{f}_t * \vec{C}_{t-1} + \vec{\imath}_t * \tilde{C}_t$ //当前时刻的单元状态
$\vec{h}_t = \vec{o}_t * \tanh(\vec{C}_t)$ //输出值
$\mathrm{F}(H_i, h_{it}) = H_i{}^t h_{it}$ //t 时刻输出值与输出向量的相似度
$a_t = \mathrm{Softmax}(\mathrm{F}(H_i, h_{it})) = \frac{\exp(\mathrm{F}(H_i, h_{it}))}{\sum_j \exp(F(H_i, h_{ij}))}$ //t 时刻的权重值
$Attention(Q, K, V) = \sum_j a_i v_i$ //最后的输出结果
endfor
return $a_1, a_2, \dots, a_T$

3.3 LSTM-Attention-News 预测模型

预测模型需要利用价格信息与新闻信息两种不同的特征，LSTM 预测模型及其预测过程与图 3-2 网络架构描述得一致。

（1）股票价格数据通过 3.2 节描述的缺失值处理和 Z-Score 归一化后，得到股票的价格特征序列。

（2）股票相关新闻文本数据，通过 3.3 届描述的朴素贝斯情感量化模型进行特征化，对当天所有新闻求得的情感值求和后，也进行 Z-Score 归一化，得到股票的新闻特征序列。

（3）将股票的前 a 个（实验中，本文设置 a 的值为 7）交易日的价格特征序列和新闻特征序列作为输入，经过 LSTM 层保持股票信息并提取其特征，获得隐藏层不同神经元的输出值。

（4）将 LSTM 层不同神经元的输出值进行相似度计算（Attention 层），通过 Softmax 函数归一化后得到重要程度的权重向量。

（5）最后权重向量经过多层感知机和 sigmoid 函数得到下一个交易日股价的涨跌概率，用$\hat{y}[0]$与$\hat{y}[1]$分别表示股价下跌与上涨的概率。最终结果取 Max($\hat{y}[0]$, $\hat{y}[1]$)所代表的含义。

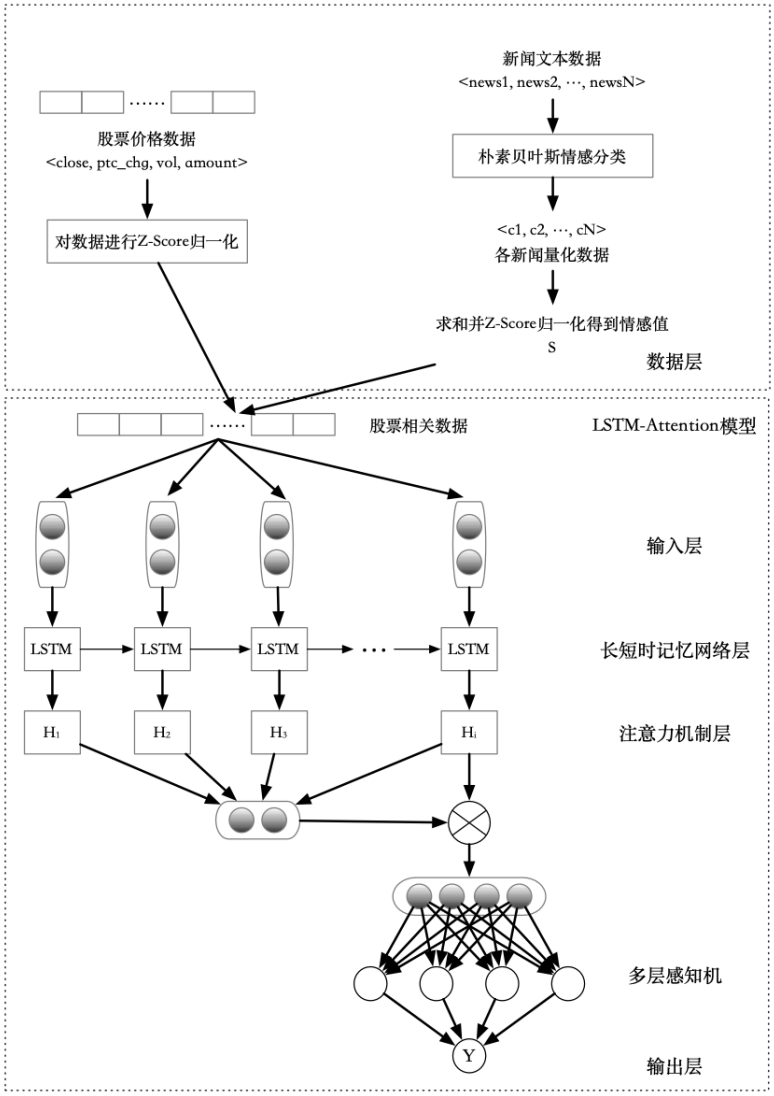


图 3-2 LSTM-Attention-News 网络架构

3.4 模型参数学习

LSTM-Attention-News 预测模型中的参数需要经过训练得到，将整个过程中参数集合用符号θ表示。为了学习参数θ，需要定义模型的损失函数。其中，y 为真实样本标签假设在一段长为 T 的交易日内，观测到某只股票的价格序列为$\vec{p} = p_1, p_2, \dots, p_T$，相关新闻序列为$\vec{d} = d_1, d_2, \dots, d_T$。经过第三章股票数据的特征化之后，获得归一化后的价格序列$\vec{q} = q_1, q_2, \dots, q_T$和新闻情感序列$\vec{s} = s_1, s_2, \dots, s_T$，结合后的特征序列$\vec{x} = x_1, x_2, \dots, x_T$通过 LSTM-Attention-News 模型整合后，输出一个结果向量，即 h_{T+1}。如式 3-11 所示：

$$h_T = lstm\_attention(x_1, x_2, \dots, x_T) \quad \text{（式 3-12）}$$

因为模型最终的目的是预测股票的涨跌，本质上是一个二分类的问题，所以适合使用交叉熵损失函数，如式 3-13 所示。其中 y 为真实样本标签，即第 t+1 天的股价涨跌的情况。

$$L(h; \theta) = -[y \log \hat{y} + (1 - y) \log (1 - \hat{y})] \quad \text{（式 3-13）}$$

模型的参数学习使用了 RMSProp 算法。该算法是对随机梯度下降（SGD）算法的一种改进，它能缓冲 SGD 算法中梯度的剧烈变化。LSTM-Attention 预测模型的代码实现使用了 Keras 库。

3.5 本章小结

本章详细阐述了本文提出的 LSTM-Attention-News 模型的股票预测方法。3.1 节介绍了 LSTM 层的算法流程。3.2 节引入了 Self-Attention 机制。3.3 节整体介绍了预测模型的计算流程。最后 3.4 节给出了该模型的损失函数和学习方法。

4 实验及结果分析

4.1 实验环境

本文实验所采用的的硬件环境和软件环境如下所示：

（1）PC 机：MacBook Pro (Retina, 15-inch, Mid 2014)

（2）处理器：2.5 GHz 四核 Intel Core i7

（3）内存：16GB

（4）操作系统：macOS Catalina 10.15.3

（5）开发环境：Pycharm、Jupyter notebook

4.2 数据采集

4.2.1 股票数据采集

股票交易数据采集使用第三章 3.1.1 小节介绍的数据采集方法，采集了平安银行（SZ000001）、万科 A（SZ000002）、格力电器（SZ000651）三支股票从 2017 年 1 月 1 日到 2019 年 12 月 31 日的股票行情数据，存入 MongoDB。数据表信息如表 4-1 所示。

表 4-1 股票行情数据信息表

股票	条数
平安银行	731
万科 A	726
格力电器	724

MongoDB 文档类似 JSON 对象，其中平安银行存储文档中的一条数据如下所示：

```
{
    "date":20191231,
    "open":16.57,
```

"low":16.31,
"close":16.45,
"change":-0.12,
"pct":-0.7242,
"vol":70442.25,
"amout":1154704.348
}

4.2.2 新闻数据采集

新闻数据采集使用第三章 3.1.3 小结介绍的数据采集方法。利用网页爬虫与网页分析工具，可以从 xueqiu.com 获得股票新闻相关数据。本文以平安银行（SZ000001）为例，介绍新闻数据采集过程。

雪球网通过在网页地址后加入股票代号来区分不同的股票。平安银行的相关资讯网页的 URL 就为 www.xueqiu.com/S/SZ000001。如图 4-1 所示，新闻消息按时间顺序由近及远排列，每则新闻的格式相同，新闻与新闻之间通过分栏符相隔。一则新闻由时间、标题、正文和消息来源 URL 构成。时间位于消息的上方，以灰色消耗字体显示，表明了这则消息发布的具体时间；标题在时间下面，以黑泽搭好字体显示，言简意赅表明了新闻的主要内容；新闻的正文位于消息的中间，以黑色小号字体显示，相对于标题内容更加详实。

主要抓取步骤如下：

（1）连接目标网页。程序使用 Selenium 库，启动一个模拟浏览器，程序模拟浏览器操作，打开雪球网中平安银行股票的网页。这时，浏览器会与雪球网的服务器完成通信，获得需要的数据。浏览器页面此时会显示平安银行股票的相关资讯。

（2）获取页面代码。模拟浏览器提供了获取页面代码的接口，程序从浏览器获得目标网页的 HTML 代码，这段代码以 HTML 格式存储了网页的页面内容，其中包含该股票的新闻资讯。

图 4-1 平安银行雪球网页面

（3）解析目标网页。程序调用 Beautiful Soup 库，解析 HTML 代码的结构。利用 HTML 文本中的 tag，id 和 class 等属性，在 HTML 的 DOM Tree 中定位新闻数据。如图 2-2 所示，class 属性为“timeline_item”的类表示页面的每一则信息，该页面显示有 10 则新闻。而进一步打开该标签可以具体定位新闻的时间、标题、正文和来源在 HTML 代码中所处的位置，如图 4-3 所示。

图 4-2 平安银行新闻页面和部分 HTML 代码

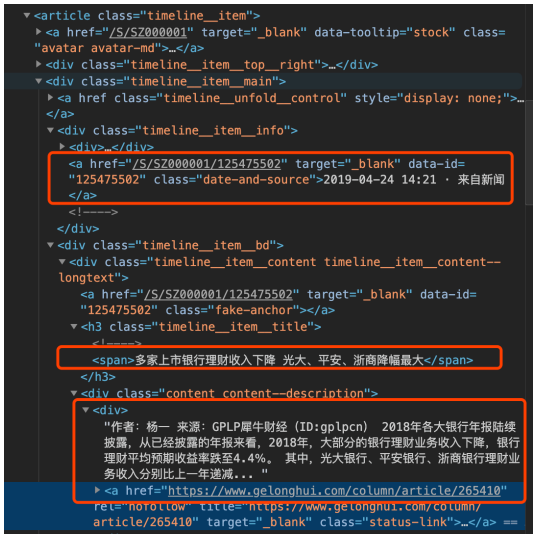


图 4-3 平安银行新闻页面部分 HTML 代码

本次试验采集了平安银行（SZ000001）、万科 A（SZ000002）、格力电器（SZ000651）三支股票从 2017 年 1 月 1 日到 2019 年 12 月 31 日的新闻数据，存入 MongoDB。数据信息如表 4-2 所示。其中平安银行某几天的数据如下所示：

{
"date":20191224,
"news":[
{
"titile":"平安银行再次中标中央财政非税收入收缴代理银行资格",
"content":"近日，在财政部举行的“2019 中央财政非税收入收缴代理银行项目公开遴选”中，平安银行凭借特色的金融科技优势与全方位的服务方案，从……",
"URL":"https://finance.sina.com.cn/stock/relnews/cn/2019-12-24/doc-iihnzahi9663931.shtml"
},{…},{…}
]
}

表 4-2 股票新闻数据信息表

股票	条数
平安银行	3153
万科 A	8739
格力电器	8625

4.3 数据处理

4.3.1 股票数据处理

为了提高模型的收敛速度和提高计算精度，本实验采用第 3 章的 3.2.2 小节描述的 Z-Score 归一化方法来对股票的收盘价、涨跌幅、成交量和成交金额进行归一化处理。并且，因为收盘价、涨跌幅、成交量和成交金额四个特征的量纲差别较大，如果不对数据进行归一化处理，模型可能倾向于数值较大的特征而忽略数值较小的特征，所以归一化操作也可以提高模型预测的准确率。

4.3.2 股票新闻处理

本实验采用第 3 章 3.3 节描述的基于朴素贝叶斯量化新闻情感的方法来对股票新闻数据进行处理。将新闻这类文本数据转为为 c 值，c 为-1、1 值代表新闻对该股票利空和利好两类情感。

量化模型训练的数据来源于：

https://github.com/wwwxmu/Dataset-of-financial-news-sentiment-classification

发布的人工打标的数据集。数据集为雪球网上万的资讯发布的正负面新闻标题，数据集包含 17149 条新闻数据，包括日期、公司、代码、正/负面、标题、正文 6 个字段，其中正面新闻 12514 条，负面新闻 4635 条。如图 4-4 所示。训练量化模型使用数据集中 4500 条正面新闻和 4500 条负面新闻。其中 70%的数据作为 train 样本，30%的数据作为 test 样本。

日期	公司	代码	正负面	标题	正文
2018年3月15日	平安银行	1	1	A股上市银行	A股上市银行首份年报出炉，平安银行(000001)3月14日晚披露年报，公司2017年营收为1057.86亿元，同比下降1.79%；净利为231.89亿元，同比增长2
2017年7月27日	万科A	2	1	万科A：昆明	万科A：昆明万科将为昆明抚仙湖项目借款提供担
2017年11月8日	万科A	2	1	万科A：截至	万科A：截至10月末借款余额为1673亿元
2018年4月26日	万科A	2	1	万科A：一季	证券时报网（www.stcn.com）04月25日讯万科A(行情000002,诊股)发布一季报称，报告期内，集团实现营业收入308.3亿元，同比增长65.8%；实现净利
2019年1月10日	万科A	2	1	万科A：广信	万科A：广信房产资产包完成交割
2018年4月9日	万科A	2	1	万科A：201	原标题：万科A：2018年前三月实现合同销售金额1542.6亿元　　中国网财经4月8日讯 4月8日，万科A发布三月份销售简报显示，2018年3月份公司实
2018年4月26日	万科A	2	1	万科A：一季	新浪财经讯 4月25日消息，发布一季报称，期内实现营业收入308.3亿元，同比增长65.8%；净利润8.9亿元，同比增长28.7%。
2017年3月10日	万科A	2	1	万科A 3月24	万科A 3月24日召开董事会
2018年3月8日	深振业A	6	0	深振业A遭证	3月8日，在深圳颇具知名度、曾被评为“深圳十强”的上市公司(000006.SZ)停牌半年之后终于复牌，但在二级市场上遭到投资者“用脚投票”。
2017年7月17日	全新好	7	1	全新好上半	全新好上半年净利预计翻倍
2018年2月28日	全新好	7	1	全新好第三	北京商报讯（记者 崔启斌 马换换）在今年2月份汉富控股有限公司（以下简称“汉富控股”）曾斥资近10亿元接盘了全新好（000007）大股东的所有
2018年5月18日	神州高铁	8	1	神州高铁：：	与腾讯云计算签署框架合作协议，共同推动移动互联网、云计算、物联网、大数据、人工智能等互联网技术在轨道交通产业领域的应用落地。
2018年5月30日	美丽生态	10	0	【美丽生态	【美丽生态：涉嫌信披违法违规 公司及实控人遭证监会立案调查】美丽生态(000010)5月29日晚公告，公司、公司控股股东五岳乾坤、公司董事丁熊秀
2018年5月30日	美丽生态	10	0	【美丽生态	【美丽生态：涉嫌信披违法违规 公司及实控人遭证监会立案调查】美丽生态(000010)5月29日晚公告，公司、公司控股股东五岳乾坤、公司董事丁熊秀
2018年5月30日	美丽生态	10	0	美丽生态：：	e公司讯，美丽生态(000010)5月29日晚公告，公司、公司控股股东五岳乾坤、公司董事丁熊秀、公司实际控制人蒋文和深圳市盛世泰富园林投资有限公
2018年5月30日	美丽生态	10	0	美丽生态：：	e公司讯，美丽生态(000010)5月29日晚公告，公司、公司控股股东五岳乾坤、公司董事丁熊秀、公司实际控制人蒋文和深圳市盛世泰富园林投资有限公
2018年5月30日	美丽生态	10	0	美丽生态：：	美丽生态(行情000010,诊股)5月29日晚公告，公司、公司控股股东五岳乾坤、公司董事丁熊秀、公司实际控制人蒋文和深圳市盛世泰富园林投资有限公
2018年12月13日	深物业A	11	1	深物业A：拟	e公司讯，(000011)12月12日晚公告，公司与控股股东深投控签订股权转让意向书，拟以现金方式或双方认可的其他合法方式收购投控物业100%股权。
2018年12月13日	深物业A	11	1	深物业A：拟	深物业A：拟收购控股股东旗下投控物业100%股权
2019年1月30日	神州长城	18	0	神州长城：：	e公司讯，神州长城(行情000018,诊股)1月29日晚发布业绩预告，预计2018年净利亏损13亿元-14.5亿元，上年同期盈利3.8亿元。2018年，因银行等金融

图 4-4 训练集样本数据

4.3.2.1 生成特征词库

基于朴素贝叶斯算法生成特征词库的具体流程如下所示。

（1）导入 train 样本，统计正面新闻和负面新闻的先验概率；

（2）读取新闻的标题和正文，使用 pkuseg 的新闻领域模式进行中文分词；

（3）匹配由清华大学自然语言处理与社会人文计算实验室整理的 THUOCL 财经词汇，得到财经样本词汇；

（4）使用 TF-IDF 方法对财经样本词汇计算权重提取特征词；

（5）统计特征词的类条件概率，得到特征词库。

4.3.2.2 模型评价指标

实验常用的的评价指标有以下四种，分别是准确率、精确率、召回率、F1 值。

（1）Accuracy（准确率）是指在分类结果中，正确预测的数量与样本总数的比值。

（2）Precision（精确率）是指在分类结果中，正确预测的正类数与预测为正类的样本数的比值。

（3）Recall（召回率）是指在分类结果中，正确预测的正类数与实际为正类的样本数的比值。

（4）当正负样本不均衡时，准确率不适用于作为评价指标，而精确率与召回率精确率和召回率通常是此消彼长的，很难兼得。所以需要用 F1（F1-score）指标来综合考虑，其定义如式 4-1 所示。

$$F1 = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad \text{（式 4-1）}$$

本次实验的 test 样本为 1350 条正面新闻和 1350 条负面新闻，测试结果的混淆矩阵如表 4-3 所示。

表 4-3 量化模型测试结果的混淆矩阵

	实际为正面新闻	实际为负面新闻
识别为正面新闻	1293	101
识别为负面新闻	57	1249

准确率为：

$$Accuracy = \frac{1293 + 1249}{1350 + 1350} = 94.19\%$$

精确率为：

$$Precision = \frac{1293}{1293 + 101} = 92.55\%$$

召回率为：

$$recall = \frac{1293}{1293 + 57} = 95.78\%$$

F1 值为：

$$F1 = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} = 94.24\%$$

从实验结果可以看出，该分类结果的 F1 值达到了 94.24%，说明本次实验中朴素贝叶斯分类器的效果优良，基本能将新闻文本的情感正确分类，可以将每天的新闻文本用该方法进行情感量化，并作为股票预测模型的输入。

4.3.2.3 新闻数据量化

对爬取到的平安银行（SZ000001）、万科 A（SZ000002）、格力电器（SZ000651）三支股票的新闻数据使用朴素贝叶斯分类模型进行情感分类，最后按日期将所有新闻的类别值 c 进行求和得到某股票当天的新闻情感量化指标 s。

4.4 股票预测

股票涨跌预测实验中，本文将每只股票的数据（价格归一化数据和相关新闻量化数据）按时间顺序排列。将前 70%的数据作为训练集，后 30%的数据作为测试集。以测试集上涨跌预测的准确率 Accuracy 值作为模型优劣的标准。准确率是指在分类结果中，正确预测的数量与样本总数的比值。

其中万科 A（SZ000002）股票的训练过程由图 4-5 所示，可以看出在 75 个 epoch 后，训练集和测试集基本都收敛，并且模型的表现相似，模型训练效果达到预期。

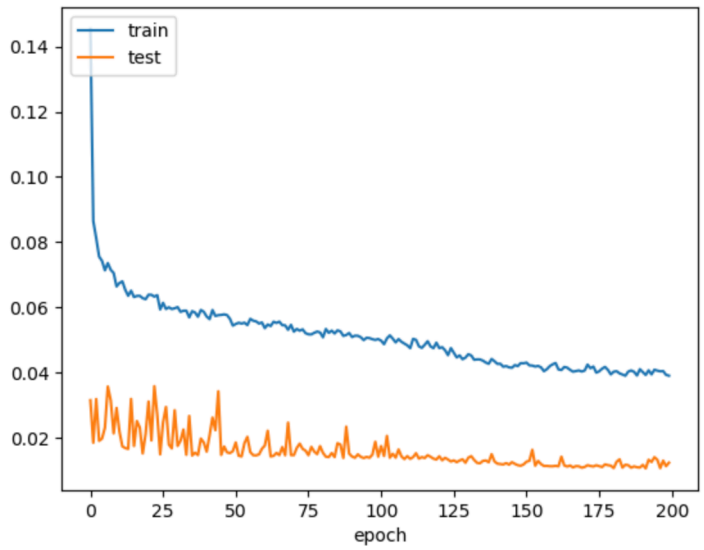


图 4-5 LSTM-Attention 网络的训练过程

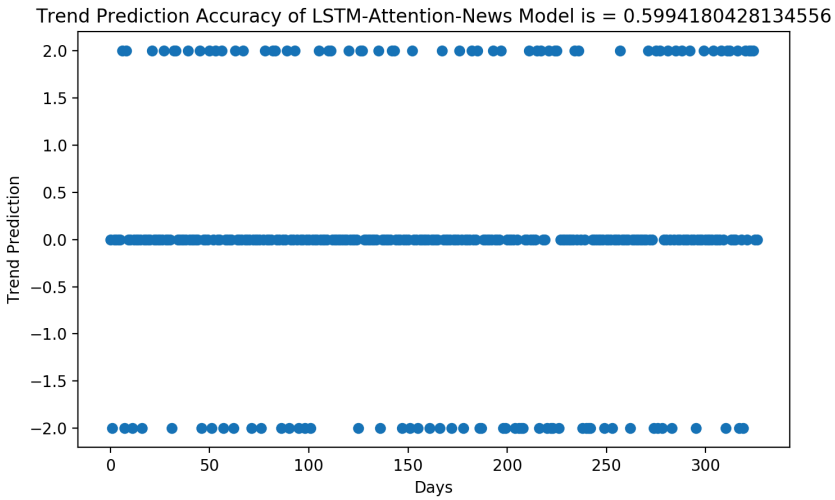


图 4-6 预测准确度表示图

预测结果的输出值为-1，1，分别表示预测该股票当天收盘价的下跌和上涨。万科A（SZ000002）股票的预测准确度如图 4-6 所示，横坐标 Days 表示预测的天数；纵坐标 Trend Prediction 表示预测准确度，含义为预测涨跌值-真实涨跌值，故当值为 0 时，表示预测正确。由图可以看出，整体预测正确的分布较为均匀，最终预测的准确率约为 59.94%，说明该模型在股票的涨跌预测上有一定的作用。

表 4-4 模型测试结果准确率比较

股票代码	SVM	SVM-News	LSTM	LSTM-News	LSTM-Attention	LSTM-Attention-News
SZ000001	0.5271	0.5512	0.5431	0.5635	0.5735	0.5801
SZ000002	0.5433	0.5439	0.5463	0.5532	0.5788	0.5994
SZ000651	0.5126	0.5487	0.5377	0.5789	0.5742	0.5925
平均	0.5277	0.5479	0.5423	0.5652	0.5755	0.5907

本文使用文献 12 提出的 SVM 模型和文献 29 提出的 LSTM 模型，与论文中的模型进行比较，并且对各模型使用不同的数据集，以此来探究新闻情感特征对预测效果的影响。共有六种实验方案，其在平安银行（SZ000001）、万科 A（SZ000002）、格力电器（SZ000651）三支股票上的预测准确率结果如表 4-4 所示。

（1）第一种方法记为 SVM，以前 3 个交易日的股票价格数据序列为特征，SVM 为模型进行预测；

（2）第二种方法记为 SVM-News，以前 3 个交易日的股票价格数据序列和新闻量化数据为特征，SVM 为模型进行预测；

（3）第三种方法记为 LSTM，以前 3 个交易日的股票价格数据序列为特征，LSTM 为模型进行预测；

（4）第四种方法记为 LSTM-News，以前 3 个交易日的股票价格数据序列和新闻量化数据为特征，LSTM 为模型进行预测；

（5）第五种方法记为 LSTM-Attention，以前 3 个交易日的股票价格数据序列为特征，LSTM-Attention 模型进行预测；

（6）第六种方法记为 LSTM-Attention-News，为本文第 4 章提出的预测方法，以前 3 个交易日的股票价格数据序列和新闻量化数据为特征。

4.5 结果分析

根据实验结果，本文有如下解释和分析：

（1）所有模型的预测准确率均在 51%-60%之间，最高能达到 59.94%。这个结果可以被半强式有效市场理论解释。该理论认为，所有公开信息相互影响形成了当前股票价格，也就是说，无论基本面方法或技术分析都无法取得短期（一天或一周）的超额投资回报。尽管如此，但本文相信个别股票的数据能提供未来价格动向（长期）的某种暗示，这也是本文研究工作的意义。

（2）在相同数据输入的情况下，LSTM-Attention 模型的结果几乎都比传统的 SVM 分类模型和单一的 LSTM 模型要好。加入了 Attention 机制的 LSTM 网络对于股票走势预测这类问题能够综合历史的数据，从而表达出更多的信息。

（3）加入了新闻量化后的特征对于三个模型的预测平均效果有约 2%左右的提升，说明对应股票的新闻情感信息在一定程度上可以表达出市场对该股票的预期，更有助于预测股票涨跌。

（4）从实验结果看本文提出的预测模型能更好得结合新闻情感量化特征在预测结果上达到 59.07%的准确率。

4.6 本章小结

本章 4.1 节介绍了本次实验所用的软硬件环境。4.2 节介绍了股票价格和股票新闻两种数据的采集过程，利用爬虫工具和页面解析工具抓取了三只股票的相关数据。4.3 节讲解了数据的处理过程，分析了新闻量化模型的分类结果的 F1 值达到 94.24%，可以对新闻文本进行情感分类。在 4.4 节和 4.5 节对三只股票进行了不同模型的预测和分析，表明本文提出的方法与对照组有 2%到 5%的提升。

5 总结与展望

5.1 论文工作总结

股票市场在我国整体经济中占据重要地位，整个股市大盘的起伏侧面反映了国家的经济状况。而对于股票投资者和投资机构来说，如果能知道未来股票的大致走势，就能为其将来的投资策略提供非常有价值的参考意义。本文主要研究的课题是基于新闻情感量化和 LSTM 网络的股票预测模型的设计与实现。针对以下几个方面展开了研究工作：

（1）实现了股票每日数据和其相关新闻文本的采集，使用 Tushare 接口和 Selenium、Beautiful Soup 工具采集到股票行情数据和相关新闻文本，能够为后续的股票预测模型提供真实可靠的数据。

（2）提出了基于朴素贝叶斯算法的新闻文本情感量化模型，使用基于新闻领域的 pkuseg 工具进行准确地中文分词和匹配 THUOCL 财经词汇，通过朴素贝叶斯算法对新闻文本进行情感分类，准确率达到 94.24%，表明该量化模型可以为股票预测模型提供可靠的股票相关财经新闻的情感数据。

（3）提出了基于 LSTM-Attention 网络的股票预测模型。在长短记忆神经网络 LSTM 中引入了 Attention 注意力机制，来学习不同时刻信息之间的相互影响和相互作用，从而改善模型预测的效果。探究提取新闻事件的情感信息作为输入源之一对股票收盘价的涨跌走势预测带来的影响

5.2 下一步工作展望

股票价格涨跌预测涵盖了经济学、数学和计算机科学等多个领域，是一个非常复杂的交叉学科问题。由于股票价格涨跌变化有着众多的影响因素，现实生活中股票价格也十分多变，目前已提出的模型尚不能完全囊括所有影响股票价格涨跌趋势的因素。本课题探究了将股票相关的新闻情感信息融入预测方法，提出的基于新闻情感量化和

LSTM-Attention 网络的股票预测模型，也仍存在着某些不足，未来的改进方向可以在本文的基础上从以下几个方面展开：

（1）在数据采集方面仅采集了三只股票数据以及雪球网的新闻数据，未来可以采集更多的股票数据记以及各类专业的股票文本数据，不仅仅局限于新闻，如国家政策发布、公司财报发布和社会舆论等。

（2）本文提出的新闻情感特征是通过本文分类的方式来进行提取的，虽然有一定的效果，但是在量化过程中丢弃了诸多的信息，只能粗略地表示一篇新闻的利好利空性，不能对其情感的程度进行更准确的量化。未来可结合 NLP 领域的相关技术对文本表达的情感进行更准确的量化以及对不同事件进行分级。

（3）本课题设计的模型仅仅预测了下一交易日特定股票的涨跌趋势，没有对涨跌的幅度进行考虑和最后的收益进行考虑。未来研究可模拟实际交易策略，判别模型是否能产生良性收益。

致谢

两年的岁月转瞬即逝，再回首，浸透着汗水与欢笑的学生生涯已经进入了尾声。研究生的两年在我 18 年的学生生涯中看似占比不大，却对我产生了举足轻重的影响。无论是导师鲁宏伟教授在论文方面对我的悉心指导，还是甘早斌副教授在平时学术生活中的谆谆教诲，亦或是同门和师兄师姐在学习之余对我的照顾与引导，无不让我对在华中科技大学的生活充满着回忆和不舍。

对父母，我感恩。从大学到研究生再到收获去深圳工作的机会，我一直没有待在父母身边好好陪伴他们。对此，他们从不抱怨，他们永远以最无私的心尊重着我的每一个决定，只希望我健康快乐。正是这份看似平凡的退让，使得我在每一个人生节点上，都能无所顾忌的做出自己最无悔的选择。

对老师，我感激。鲁老师和甘老师自己的学术研究也十分繁忙，但是这从未成为老师们拒绝我求知、求学的理由。相反，在论文写作过程中，大到研究方向和论文框架制定，小到论文格式、标点符号，两位老师都给予了我相应的指点。一次又一次发问，一次又一次的解答，是老师的耐心和细心，让我去除浮躁更沉下心来完成论文的写作。

对同学，我感谢。入学伊始，新的学校新的实验室，我也曾感到陌生和无措。是白爱师兄、汤景仁师兄和张晨师兄热情相迎，让我快速适应了华科的学习氛围。当学习、实习、求职的压力同时袭来时，是蔡媛欢、魏宜、许伦祥和喻宗尧等同学一起互相鼓励、分享经验，让我找准方向集中发力，直面秋招浪潮。

在华科的日子，我像是广袤森林中的一颗期待成长的树苗。此时，老师如参天大树，予我荫庇，以他直冲天空的背脊教我未来要成长成怎样的模样；同学如林中花朵，予我芬芳，以其可爱面庞让我的生活去单调而只留美好；学校如森林，予我空间，让我心无旁骛的学习毫无杂念的成长。我真心诚意的对华科的一切说一句谢谢，愿未来的自己，以梦为码、继续拼搏。

参考文献

[1] Fama E F. The behavior of stock-market prices. The journal of Business, 1965, 38(1): 34~105

[2] Schumaker R P, Chen H. A quantitative stock prediction system based on financial news. Information Processing & Management, 2009, 45(5): 571~583

[3] Day M Y, Lee C C. Deep learning for financial sentiment analysis on finance news providers. Advances in Social Networks Analysis and Mining. IEEE, 2016. 1127~1134

[4] 张梦吉,杜婉钰,郑楠.引入新闻短文本的个股走势预测模型.数据分析与知识发现,2019,3(05):11~18

[5] 傅魁,刘玉洁,陈美丽.基于财经新闻情感倾向值的股票价格预测.北京邮电大学学报(社会科学版),2019,21(01):87~100

[6] 杨建辉,沈淑.新闻媒体报道对股价同步性影响的实证分析.统计与决策,2018,34(12):156~159

[7] Tang D, Wei F, Yang N, et al. Learning Sentiment-Specific Word Embedding for Twitter Sentiment Classification. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. ACL, 2014.1555~1565

[8] 黄秋萍,周霞,甘宇健等.SVM 与神经网络模型在股票预测中的应用研究.微型机与应用,2015,34(05):88~90

[9] 李坤,谭梦羽.基于小波支持向量机回归的股票预测.统计与决策,2014(06):32~36

[10] Wenjuan Mei. Stock price prediction based on ARIMA-SVM model. Institute of Management Science and Industrial Engineering.Proceedings of 2018 International Conference on Big Data and Artificial Intelligence.Computer Science and Electronic Technology International Society, 2018. 55~61

[11] M. M. Gowthul Alam,S. Baulkani. Local and global characteristics-based kernel hybridization to increase optimal support vector machine performance for stock market prediction. Knowledge and Information Systems,2019,60(2):971-1000

[12] 邬春学,赖靖文.基于 SVM 及股价趋势的股票预测方法研究.软件导刊,2018,17(04):42~44

[13] 黄同愿,陈芳芳.基于 SVM 股票价格预测的核函数应用研究.重庆理工大学学报(自然科学),2016,30(02):89~94

[14] Chien-Feng Huang，A hybrid stock election model using genetic algorithms and support vector regression. Applied Soft Computing, 2012(12) : 807~818

[15] 李辉,赵玉涵.基于 DFS-BPSO-SVM 的股票趋势预测方法.软件导刊,2017,16(12):147~151

[16] Feng Zhou,Qun Zhang,Didier, et al. Cascading logistic regression onto gradient boosted decision trees for forecasting and trading stock indices. Applied Soft Computing Journal,2019,84:105747

[17] 王领,胡扬.基于 C4.5 决策树的股票数据挖掘.计算机与现代化,2015(10):21~24

[18] 王禹,陈德运,唐远新.基于 Cart 决策树与 boosting 方法的股票预测.哈尔滨理工大学学报,2019,24(06):98~103

[19] 陈宇韶,唐振军,罗扬等.皮尔森优化结合 Xgboost 算法的股价预测研究.信息技术,2018(09):84~89

[20] 王燕,郭元凯.改进的 XGBoost 模型在股票预测中的应用.计算机工程与应用,2019,55(20):202~207

[21] Zhang Guoying, Chen Ping. Forecast of yearly stock returns based on Adaboost integration algorithm. 2017 IEEE International Conference on Smart Cloud. SmartCloud, 2017. 263~267

[22] Xiao-dan Zhang,Ang Li,Ran Pan. Stock trend prediction based on a new status box method and AdaBoost probabilistic support vector machine. Applied Soft Computing, 2016, 49:385~398

[23] 张贵勇. 改进的卷积神经网络在金融预测中的应用研究[硕士学位论文]。郑州大学，2016.

[24] 胡悦.基于卷积神经网络的股票市场择时模型——以上证综指为例.金融经济,2018(04):71~74

[25] Hakan Gunduz, Yusuf Yaslan, Zehra Cataltepe. Intraday prediction of Borsa Istanbul using convolutional neural networks and feature correlations. Knowledge-Based Systems,2017,137:138~148.

[26] 陈祥一. 基于卷积神经网络的沪深 300 指数预测[硕士论文]。北京邮电大学，2018

[27] Stoean Catalin, Paja Wiesław, Stoean Ruxandra, et al. Deep architectures for long-term stock price prediction with a heuristic-based strategy for trading simulations.. PloS one,2019,14(10)

[28] Kim Taewook,Kim Ha Young. Forecasting stock prices with a feature fusion LSTM-CNN model using different representations of the same data.. PloS one,2019,14(2)

[29] 彭燕,刘宇红,张荣芬.基于 LSTM 的股票价格预测建模与分析.计算机工程与应用,2019,55(11):209~212

[30] 陈佳,刘冬雪,武大硕.基于特征选取与 LSTM 模型的股指预测方法研究.计算机工程与应用,2019,55(06):108~112

[31] 史建楠,邹俊忠,张见等.基于 DMD-LSTM 模型的股票价格时间序列预测研究.计算机应用研究,2020,37(03):662~666

[32] 曾安,聂文俊.基于深度双向 LSTM 的股票推荐系统.计算机科学,2019,46(10):84~89

[33] 王子玥,谢维波,李斌.变步长 BLSTM 集成学习股票预测.华侨大学学报(自然科学版),2019,40(02):269~276

[34] A. Jayanth Balaji, D.S. Harish Ram, Binoy B. Nair. Applicability of Deep Learning Models for Stock Price Forecasting An Empirical Study on BANKEX Data. Procedia Computer Science,2018,143:947-953

[35] Akita R, Yoshihara A, Matsubara T, et al. Deep learning for stock prediction using numerical and textual information. 2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS). IEEE, 2016. 1~6

[36] 黄丽明,陈维政,闫宏飞等.基于循环神经网络和深度学习的股票预测方法.广西师范大学学报(自然科学版),2019,37(01):13~22

[37] 朱梦珺,蒋洪迅,许伟.基于金融微博情感与传播效果的股票价格预测.山东大学学报(理学版),2016,51(11):13~25

[38] 黄润鹏,左文明,毕凌燕.基于微博情绪信息的股票市场预测.管理工程学报,2015,29(01):47~52+215

[39] 董理,王中卿,熊德意.基于文本信息的股票指数预测.北京大学学报(自然科学版),2017,53(02):273~278

[40] Misuk Kim,Eunjeong Lucy Park,Sungzoon Cho. Stock price prediction through sentiment analysis of corporate disclosures using distributed representation. Intelligent Data Analysis,2018,22(6): 1395~1413

[41] 张扬,崔晨阳.基于朴素贝叶斯模型的一种网络负面信息预警策略研究.图书馆杂志,2014,33(08):78~82