西南财经大学

Southwestern University of Finance and Economics

课程论文

学年学期：2020-2021 学年第二学期

课程名称：数据分析

论文题目：基于多源异构数据的股价预测

学生学号：41811025、41823019、41823076、41811085

学生姓名：宋佳、贾婕、王欣眉、朱颖

学　　院：经济信息工程学院

年级专业：2018 级信息管理与信息系统

（金融智能与信息管理实验班）

评语：

得　分：

评阅教师签字：

年　月

分工明细

注：本组成员有宋佳、贾婕、王欣眉、朱颖、冯云浩。在与老师讨论选题之后，冯云浩的比赛项目工作与本次选题过于重复，因此冯云浩提交自己的比赛作品，其他人继续完成本次选题。

	开题阶段	数据收集阶段	数据处理阶段	模型构建阶段	论文写作阶段
宋佳	文献调研（文本量化）	利用 scrapy 爬取新浪财经研报数据	整理文本量化方法；实现词、句、段、语法树、情感的文本特征量化	深度学习模型构建、调参、实验	深度学习模型实验部分
贾婕	文献调研（机器学习股价预测）	帮忙爬取研报数据	帮助文本量化方法 debug	机器学习线性模型构建、特征工程、调参	机器学习模型实验部分
王欣眉	文献调研（各个方面）	获取 k 线数据、整理有效的交易因子	整理文本量化方法	完成其他全部论文	
朱颖	文献调研（深度学习股价预测）	帮忙爬取研报数据	计算交易因子数据	机器学习非线性模型构建、特征工程、调参	机器学习模型实验部分
冯云浩	文献调研（各方面）	《基于多因子驱动的证券市场测度研究》			

签名：

摘要：股票价格及趋势预测是金融智能研究的热门话题。一直以来,各种各样的信息源被不断尝试用于股价预测，例如基本经济特征、技术指标、网络舆情、财务公告、财政新闻、金融研报等。然而,此类研究大多数只使用结构化数据,使用文本信息的极其少见，融合多种信息源的更为罕见。因此，本文以 A 股钢铁行业、化学原料行业以及种植业为例，尝试同时利用技术指标、结构化因子与金融研报 3 种信息源来进行股价预测。首先对不同类型的信息源数据进行针对性的处理，使其形成统一的数据集,然后深度学习模型、机器学习模型与线性模型输出表示股价涨跌振幅的判断。其中，创新实现了 CNN 与 LSTM 模型的融合，提高模型准确度。实验结果表明，MLP 神经网络的预测效果最好，研报处理出来的情感与信息熵在一定程度上对模型评估有负面作用。研究能对传统量化投资提供一定的参考。

关键词：多源异构数据，因子分析，深度学习，机器学习，量化交易策略

一、引言

随着时代的不断进步，人民生活水平日益提高。在解决温饱问题之余，有了可供投资的余财。越来越多的人将目光转向股市投资，为股市发展提供了资金条件。然而在纷繁复杂的股票市场，如何充分利用各类信息源寻找最优股成为亟待解决的问题。这不仅是投资者单方面的困惑，也是量化交易领域中学者们所关心的重点。

常用的信息源有基本经济特征、技术指标、网络舆情、财务公告、财政新闻、金融研报等。其中，随着计算机技术的进一步发展和各种信息提取和数据分析技术进一步成熟，基本经济特征、技术指标的影响力逐渐下降，基于事件驱动的股价预测模型逐渐成为了研究者和投资者的研究目标。通过对新闻内容、社交媒体情绪、金融研报信息挖掘，国内外学者采用 BP 算法、ANN 算法、人工神经网络等机器学习与深度学习算法，对股票价格变化规律的研究。但对于模型预测效果的解释，每个学者根据不同的样本数据与模型给出了不同的解释。Wiwik. Anggraeni（2017）和 Nezhad S（2016）对股票价格变化进行了长期深入的研究，并分析了 SVM、向量自回归(VAR)和差分整合移动平均自回归模型(ARIMA)的

股价预测性能，发现预测模型 SVM 比以上两者具有更好的预测效果。Cao 等对支持向量机(SVM)和 BP 神经网络进行实验比较发现 BP 神经网络的预测效果更好。上述分析表明，本文研究的关键在于建立预测模型有效检验多重信息源对股票价格预测的影响，并进行系统性检验。由此，本文提出两个研究问题：第一，增加非结构化的文本数据能否提高模型准确性，有效提高股票价格预测能力?第二，若机器学习或深度学习算法的运用能够提升模型的准确性，哪种模型效果最好?最能够预测股价趋势？本文将针对这两个问题展开具体研究。

本文的研究具有一定的理论与现实意义：传统的根据金融技术指标的基本面分析看似全面却不一定准确，对于结构复杂的大数据更是难以快速正确处理。由于人为因素影响，某些信息会被主观的进行放大或者忽略。而量化交易的工作能够处理海量的信息并且不受人为因素的影响，如量化选股模型能够在对海量数据进行分析后，得出最有可能获得较高收益承担较低风险的股票组合。本文对量化方面的研究可以弥补传统投资策略中忽略的影响因素，处理大数据并排除人为因素影响，并基于多源异构数据构建有效因子，提取出新的技术指标，带入模型进行对股票价格进行预测，从而给出投资策略。

本文第 2 二部分介绍基于各种模型与信息源的股价预测的研究现状；第三部分介绍样本数据以及针对不同信息源特征的数据处理方法；第四部分陈述多种模型的核心原理与模型设计；第五部分是实验设计及分析；最后总结全文。

二、文献综述

2.1 预测股票价格模型

近年来，大量学者开始研究如何运用机器学习方法预测股票价格或收益。李斌等(2017，2019)前后分别采用线性回归、支持向量机、神经网络、XGBoost 等机器学习算法，构建股票收益预测模型及投资组合，发现这些算法具有更高的预测准确率。具体而言，预测股票价格的模型可以分为三类：传统非线性机器学习算法、深度学习算法和传统线性机器学习模型。其中，传统线性机器学习模型大多作为参照模型来帮助学者判断其他模型预测效果情况。

在传统非线性机器学习的研究中，支持向量机(SVM)以其能很好地处理小样本、非线性、高维数问题，降低训练数据样本、防止过拟合和欠拟合等优良特性，被广泛应用于股票价格进行预测。例如，Jigar Patel（2015）等将人工神经网络(ANN)和 SVM 应用于 S&P CNX Nifty 指数预测，对比实验结果发现 SVM 预测效果优于 ANN。虽然 SVM 可以有效预测股票价格（张晨希 2006），但仍存在着难以解决多分类问题和训练大规模样本等问题。因此，王晓红、王梦瑶等人（2020）构造改进的时间相关序列（ARIMA—TGD—SVM）股票价格混合预测模型来弥补传统支持向量机（SVM）模型的不足。另外，王燕与郭元凯（2019）通过网格搜索算法对 XGBoost 模型进行参数优化构建 GS-XGBoost 的金融预测模型，并将该模型运用于股票短期预测中。

自 Lapedes（1987）等在股价预测领域应用了人工神经网络开始，许多学者开始关注深度学习在股价预测中的应用。针对股票价格的突变性、非线性和随机性，单一预测方法仅能描述股票价格片断信息等缺陷，俞国红等（2013）提出一种股票价格组合预测模型。由于 BP 神经网络暂且无法刻画复杂股市变化规律，许多研究基于 BP 神经网络预测模型做了进一步的改进，曹晓（2017）等提出了一种灰色 GARCH-BP 组合模型，这种混合模型相较于单个模型有更好预测效果；孙存浩（2020）基于 BP 神经网络模型与长短期记忆（LSTM）神经网络模型构建了 BP-LSTM 模型；赵红蕊与薛雷（2021）提出了一种在结合长短期记忆网络（LSTM）和卷积神经网络（CNN）的基础上引入注意力机制的股票预测混合模型（LSTM-CNN-CBAM）。

2.2 多源异构数据

除了上述研究之外，有学者指出模型的预测精度与参数的选择有关，参数选取越科学模型预测效果越好（Wang S.X.2016,Jinming You 等 2017），信息源越多能够提供更加丰富的信息内容和更多不同的信息层面（饶东宁，邓福栋 2017）。但是现有的股票价格预测研究主要是在数值分析框架中应用各种时间序列模型和优化方法，使用文本数据的研究相较甚少。

在数值分析框架中，大多数学者主要运用技术指标与基本面信息来构造因子组合。技术指标反映市场某一方面的信息，最常用的是相对强弱指标(RSI)、简单移动平均(MA)、指数平滑异同平均(MACD)、随机指标(KDJ)等（饶东宁等，2017）。Lee（2009）在 NASDAQ 指数上进行实验，构建了一个 SVM 模型来预测股市变化。Patel 等（2015）10 个技术指标，同时结合了 4 个印度股票价格特征，使用人工神经网络、支持向量机、随机森林(RF)、朴素贝叶斯分类(Naive Bayesian Classification)4 种不同类型的机器学习方法进行全面的对比。实验结果表明，RF 具有最好的分类性能，其次是 SVM。但随着计算机技术的进一步发展和各种信息提取和数据分析技术进一步成熟，该方法的流行程度正逐渐下降。

即利用公司财报中的数据分析公司本身价值的变化，从而预测股价的变化。由于该方法的经济学可解释性一直以来得到了经济学家以及投资者们较多的关注。例如，王亚红与程希明（2018）选取有关公司盈利能力、成长能力、营运能力、偿债能力的代表指标数据作为特征属性，建立随机森林模型并进行测试。文宇（2018）根据标的的历史价格和成交量信息扩展了更多维度的特征，并用 CNN 模型进行空间上的特征提取，对金融二级市场数据进行分析。Tsai 等（2011）选取了 19 个金融特征和 11 个经济指标作为输入变量，构造了一个集成分类预测模型。Lam（2004）使用 16 个金融变量和 11 个宏观经济变量，构建了一个基于后向传播神经网络的预测模型。然而，该方法仍存在较多不足。首先，各公司发布的财务报告格式并不统一，目前的信息提取技术较难从中获得完全准确的信息，仍需要大量时效性较低的手工标注等工作。其次，会计标准上仍有一定的模糊性，使得各公司的财务报表并不能完全横向比较。最后，由于上市公司的财务报表每年仅发布 2-4 次，时间间隔长，股票回报率存在较大风险，一旦中间出现对公司影响较大的事件，会使之前的股价预测变得不准确。目前，利用该方法进行股价预测大多是将模型预测结果与宏观交易员的经验相结合，从而避免以上提到的问题。因此该方法并不能实现完全的自动化交易。

随着人工智能尤其是自然语言处理技术的快速发展，使得人们从新闻以及事件中提取信息并理解信息的能力得到的极大提高，因此基于事件驱动的股价预测模型逐渐成为了研究者和投资者的研究目标。目前通过自然语言处理进行股价预测的任务从数据类型上可以分为两类：社交媒体内容和新闻内容。

在社交媒体内容挖掘方面，林培光与周佳倩（2020）等人爬取股民评价，提出一种基于 ConvLstm（convolutional long short term memory）的股票情感分析价格预测的深度学习模型 SCONV（semantic convolutional），发现相比一些传统模型，SCONV 在较小的样本集上可以更好地预测股票价格的走势。季子峥、沈婷婷（2020）等引入"话题-情感"等细粒度情感信息，构建了包含众多股票的长时间跨度数据集，并在此数据集上验证了细粒度情感分析对股价涨跌预测的良好效用。Bollen 等（2010）通过 twitter 对金融数据进行情感分析来研究市场预测，使用两个心情追踪的工具，通过 6 个心情维度来计算情感值。徐琳（2013）研究了网络舆情(主要是微博)对我国股票市场的影响，提出了网络舆情对股票波动可能具有正、负、超效应，并进行实验验证。胡婧（2017）等也考虑了微博舆情数据对股价变化的影响，在神经网络模型中加入微博数据来提升预测稳定性。综上，情感分析技术已广泛用于股票预测，特别是在博客方面。

在新闻信息的利用上，赵丽丽、赵茜倩等人（2012）首先采用文本挖掘技术、支持向量回归（SVR）方法将财经新闻内容量化为股市波动的一个影响因子，再利用多元回归分析方法系统地分析了互联网财经新闻信息对股市的影响。Zhai 等（2007）结合公司新闻和技术指标，利用 SVM 分类器对公司股票价格进行预测。孔翔宇与毕秀春等（2016）深度挖掘了财经新闻主题内容与股市市场的相关性，并提出了一种基于理解当日新闻主题分布来分析中国股市涨跌的 SVM 预测模型。Li（2014）等使用市场新闻和股票价格两种信源来构建预测模型，他们使用 ELM(Extreme Learning Machine)方法来挖掘隐藏数据，预测精准度与 SVM 类似，但学习速度更快。张梦吉、杜婉钰等（2019）建立基于事件的新闻分类模型，使用多输入的循环神经网络建立基于新闻事件、资金流向和公司财务的个股走势预测模型，提升股票预测准确率。结果表明新闻中国际贸易以及城市化内容与股市变动关系密切，可凭此有效预测当日股市涨跌。其他信源结合方式也都各具特色。

另外，金融研报往往也会反映投资者的推荐行为。由于市场的信息不对称性，投资者充分信赖分析师出具的预测信息，一旦研究报告透露出利好或者消极的投资风向，将会直接影响投资者的投资组合策略，当某个研报“强烈推荐”某只股票时，该股票的价格很可能上涨;当某个研报“减持”某只股票时，该股票的价格很

可能下跌。Duan（2015）等提出基于分析者的推荐行为的后验概率模型来预测股票的回报;进一步地，他们结合股票交易信息和投资者评价进行股票收益预测；特别地，使用规则提取技术建立预测系统，解决了大多数预测模型不具有解释性的问题。此外，Newman（2013）等也研究了投资者推荐行为，与大多数相关研究不同，他们使用买方数据进行分析。总之，分析师处于资金的需求端和供给端之间，他们用相对专业的金融知识和直接对企业进行调研的信息优势，第一时间了解到企业的财务和运行状况，并整理成研究报告的形式，向投资者们提供相关证券的价值信息。这个过程提高了企业信息向市场传输的速度，使企业的经营状况更快的反映到其股票价格中来，也可能对股价波动产生影响。

2.3 文献评述

目前，于事件驱动的股价预测模型的研究存在不足，尤其是如何利用股票研报进行自然语言分析建立投资策略问题。首先，传统经济方法只考虑了传统的经济金融特征与技术指标对股价的影响，这些特征一般都是数值型。其次，基于机器学习技术的各种预测方法虽然扩展了影响股价的特征，并且考虑了一些文本型数据对股价的影响，但是并没有考虑多个文本型特征共同作用时股价的变化。基于这些不足，我们在考虑影响股价的因素时，既考虑了传统的技术变量，也考虑了以研报作为来源的各种非结构文本型数据。因此，我们着重研究多信息源，以研报作为文本信息来源。

本文选取 K 线数据、交易因子、非结构化数据因子等 3 种不同的信息源来构造一个多源的股价预测模型。该预测模型使用 3 种信源融合的数据，其经过处理后形成统一的数据集，然后分别通过深度学习模型、机器学习模型与线性模型输出表示股价涨跌振幅的判断。它可以使投资者考虑更全面的信息，进行更好的投资组合选择。

三、数据选择与处理

3.1 数据描述

本文选择 A 股中钢铁行业、种植行业、化学原料行业共 86 支股票作为研究对象，包括每日每只股票的开盘价、收盘价、最高价、最低价与交易量。选择上述公司在 2020 年 1 月 1 日至 2020 年 12 月 30 日的日频数据作为样本，经数据预处理后，共得到 62780 条数据。数据来源于国泰安数据库。图 1-4 为各行业在样本区间内的价格走势图。

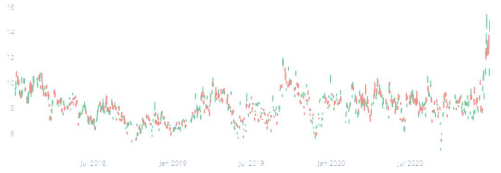

图 1 钢铁行业 K 线图

图 2 化学原料 K 线图

图 3 种植业 K 线图

图 4 A 股整体趋势图

根据同花顺 Mind Go 量化交易平台，得到技术面的结构化特征。各行业的技术指标如图 5-7 所示。根据爬取的新浪财经的股票研报，按照价值分析原则，对非结构化数据进行因子构造。因子构造结果如图 8 所示。

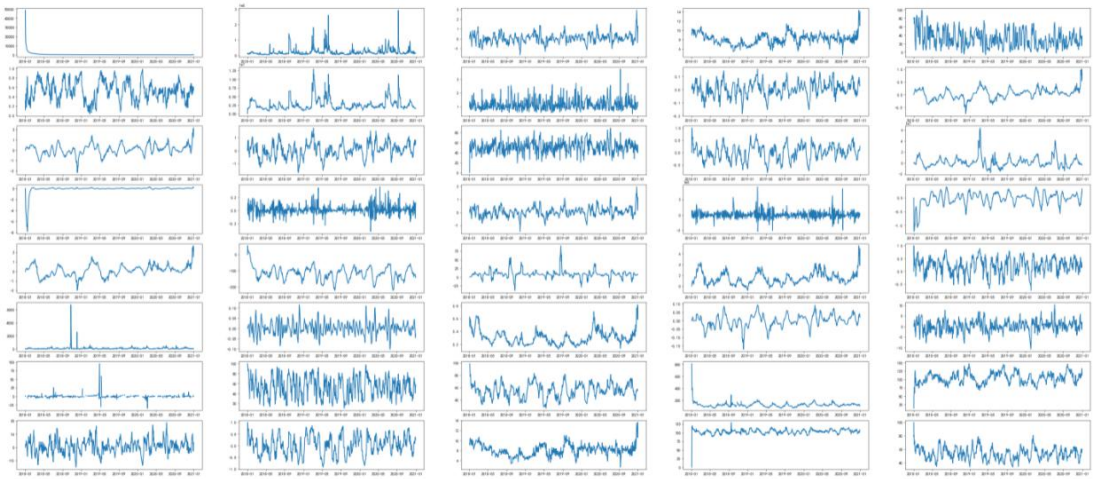

图 5 钢铁行业股市技术指标

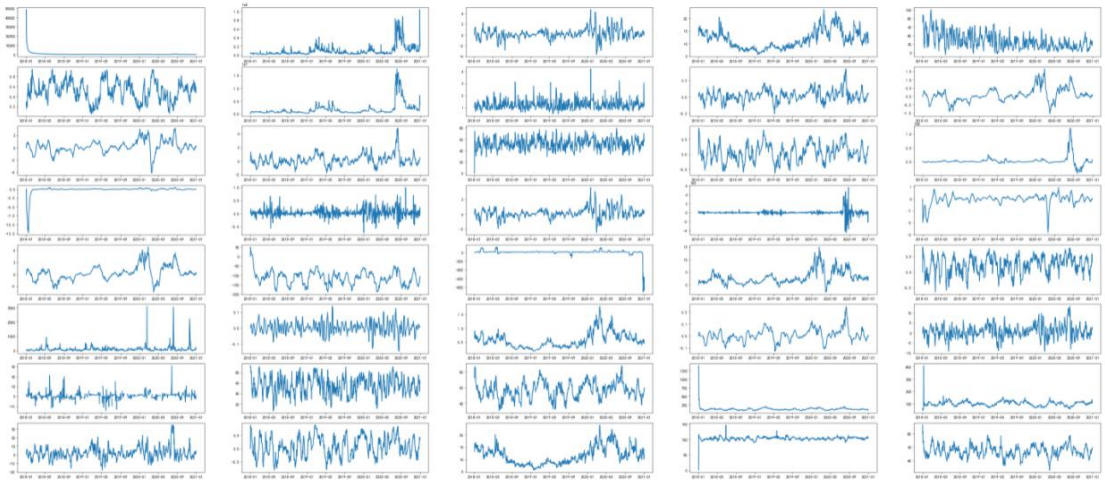

图 6 化学行业股市技术指标

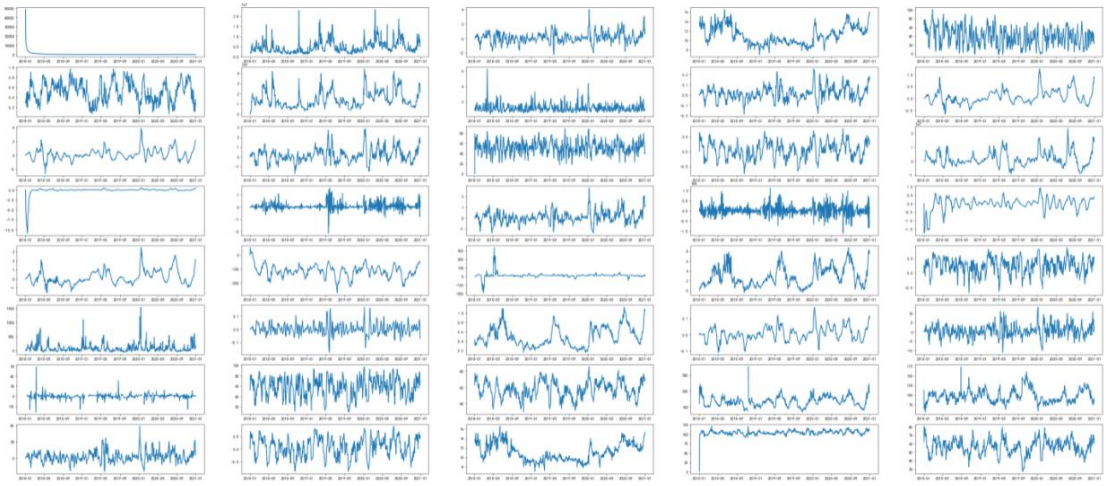

图 7 种植业股市技术指标

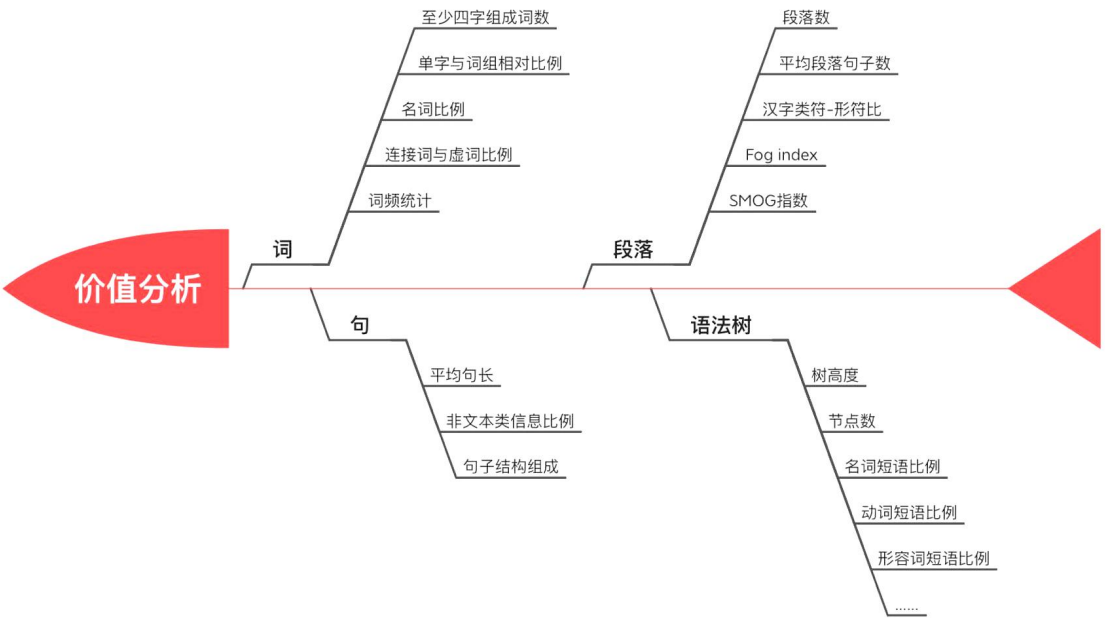

图 8 非结构化数据构造因子鱼骨图

3.2 数据预处理

数据预处理的整体流程如下所示：

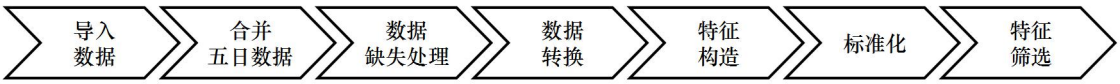

图 9 数据预处理整体流程图

导入数据后，首先按照时间顺序，对连续五天的因子数据进行合并，将数据整理为一行。最终所得到的数据因子相对于原先多了五倍，确保样本包含历史股市信息。并对 K 线数据特殊性导致因子超出阈值或者异常取 0 的情况进行多项式插值填补。再将日 K 线数据、结构化数据因子构造、非结构化数据因子三部分数据，按照日期进行拼接，转换数据类型，保证数据类型与结构正确。

针对机器学习数据，进行暴力交叉的特征组合构建，包括一阶的加减乘除，以及二阶的正态变化、平方和、平方差和绝对平方根处理。特征组合内容如图 10 所示。最终得到 6730 个特征。

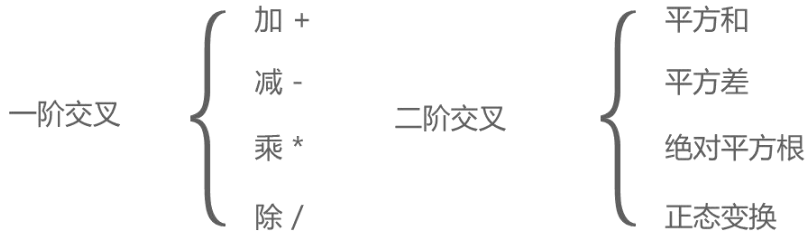

图 10 特征组合示意图

为使模型拟合过程中损失函数在各个方向上下降速度一致，对数据特征进行归一化处理。计算公式为：

$$\frac{x_i - x_{\min}}{x_{\max} - x_{\min}}$$

其中，x_{min}与x_{max}为原始数据集的最小值与最大值。经过归一化处理后，数据整体落在[0,1]范围内。

特征过多会增加模型的复杂程度，因此需进行特征选择。在结构化数据因子构造部分，通过因子回测的方法，进行因子有效性检测，选择出有效的 40 个因子，如表 1 所示。而对于非结构化数据因子与机器学习特征组合出的新特征，本文主要采用过滤式和嵌入法进行特征筛选，利用 lightGBM 的 GBDT 树集成模型

进行特征重要程度比较，其结果如图 11 所示。比较发现非结构化数据因子在机器学习模型中占有重要程度，最终保留 16 个非结构因子、2566 个特征。非结构化因子如表 2 所示。

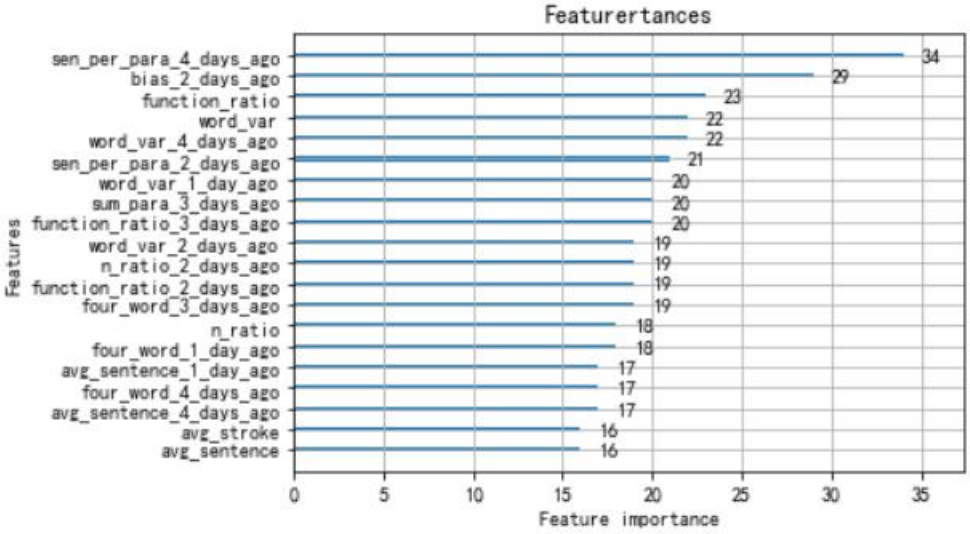

图 11 特征重要程度

表 1 结构化数据构造因子名称释义表

结构化因子符号	结构化名称	含义
CCI	顺势指标	测量股价是否已超出常态分布范围。
TAPI	指数点成交值	探讨每日成交量值与指数间的关系。
MTM	动量指标	研究股价波动的中短期技术分析工具。
VMA	变异平均线	每日开盘价、收盘价、最高价和最低价得到的数据计算平均线。
KDJ	随机指标	股票分析的统计体系。
OBV	能量潮	通过统计成交量变动的趋势来推测股价趋势。

表 2 非结构化数据构造因子名称释义表

非结构化数据因子名称	因子释义
avg_sentence	平均句长
avg_stroke	平均笔画数
four_word	至少四字组成词数

function_ratio	连接词数与虚词数
n_ratio	名词比例
nine_stroke	至少九笔字数
no_text	非文本信息比例
sen_per_para	平均段落句子数
sum_para	段落数
word_max	每句词最大频率
word_mean	每句词平均频率
word_phrase	单字与词组比例
word_var	每句词频率方差
quality_val	信息熵
pos_num	积极情感词数
neg_num	消极情感词数

四、股票价格预测模型与方法

4.1 深度学习模型

4.1.1 卷积神经网络预测模型

卷积神经网络（CNN）是一种具有局部连接、权重共享等特性的深层前馈神经网络，克服了全连接前馈网络参数太多和无法提取局部不变性特征的问题。目前的卷积神经网络一般由卷积层（Convolutional layer）、池化层(Pooling layer)和全连接层(Fully connected layer)交叉堆叠而成的前馈神经网络，使用反向传播算法进行训练。

卷积神经网络具有强大的特征提取和识别能力，在图像数据和时间序列数据的分类任务中得到了成功的应用。传统卷积神经网络的输入数据的行数和列数通常相等。而本文处理得到的股票时序数据显然难以满足此条件，为了适应股票时序数据的形式，参照 Lee 等(2017)的研究，对传统 CNN 的结构进行了调整。选用 ReLU 函数作为激活函数，在池化层选用常用的最大值池化（Maximum Pooling）进行运算；在全连接层，将最后一个全连接层前的激活函数取消，变为直接进行线性连接，得到股票价格预测值。具体来看，本文把时间序列数据视为二维数据，长为天数，宽为特征数，输入通道为 1，输出通道为 32，卷积核大小为 3。由于使用一维卷积，因此卷积核实际大小为特征数×3。在填充后的数据上进行卷积后，输出 32 个 1×5 的张量。对这 32 个张量进行首尾拼接后再进行全

连接，得到预测结果。

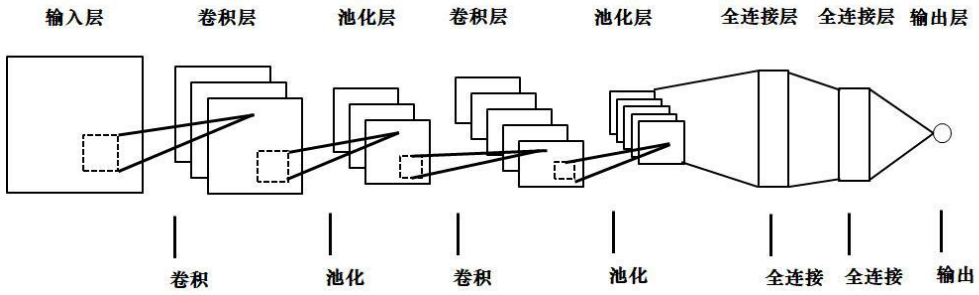

图 12 CNN 模型示意图

4.1.2 LSTM 模型

为解决循环神经网络模型由于输入序列过长，而产生的梯度爆炸和消失问题，LSTM（Long Short-Term Memory，长短期记忆网络）在 1997 年首先被 Sepp Hochreiter 和 Jurgen Schmidhuber 提出，它在 RNN 的基础上在每个神经元内加入了输入门、输出门、遗忘门三个门来控制信息的流入、存储和流出。三个门的激活函数均为 Sigmoid。输入门用来控制当前时刻神经单元的输入信息，遗忘门用来控制上一个时刻神经单元中存储的历史，输出门用来控制当前时刻神经单元的输出信息。

股票市场上的数据多为时间序列数据，数据之间的时序性无法被 BP 神经网络捕捉学习到。递归神经网络不同于传统的 BP 神经网络的结构只在层与层之间建立了连接，它在同一层的不同神经元之间也建立了连接，这样的结构使得 RNN 可以处理序列变化的数据。本文基于原数据特征情况，使用一层 LSTM 的模型结构进行预测，设置隐藏层大小为 64。

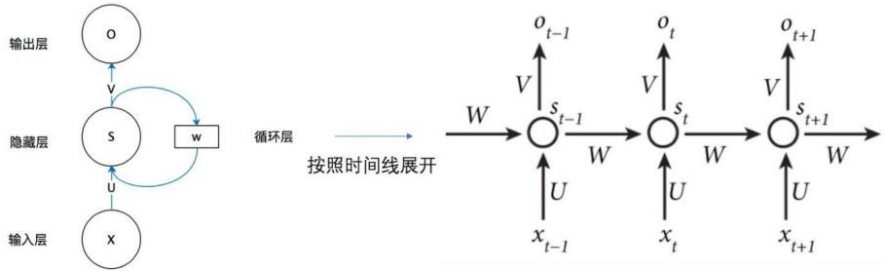

图 13 LSTM 模型示意图

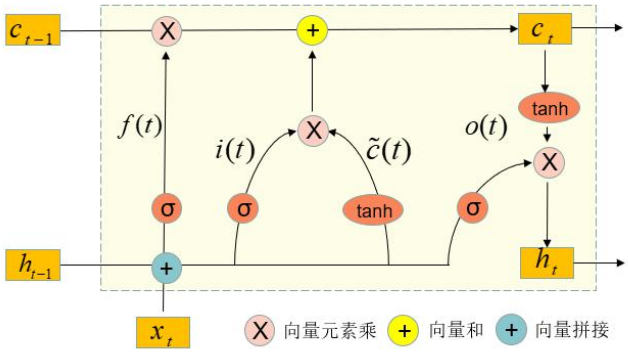

图 14 LSTM 循环单元结构

4.1.3 CNN-LSTM 模型

为了更好地应用深度学习方法对股票价格进行预测，本文利用同时提取卷积神经网络与长短期记忆神经网络的特征优势，将这两种深度学习模型融合进行预测，以更好地处理和分析金融时间序列数据。

CNN-LSTM 模型进行股票指数预测的过程可划分为两个阶段。第一阶段为 CNN 模型的特征提取阶段，第二阶段为 LSTM 模型根据时序数据进行预测阶段。

由于 CNN 中全连接层之前的作用是提取特征，全连接层的作用是分类。全连接过程相当于用全局卷积来实现这一过程，卷积的重要作用是把分布式特征映射到样本标记空间，即把特征整合到一起，输出为一个值。为了将两种深度学习模型融合到一起，本文对 CNN 模型进行修改，保留卷积层、非线性层进行特征提取，删去最后的将特征整合到一起的全连接层。从而能够利用卷积神经网络的优势提取特征。将提取出的保留时序关系的特征作为输入数据，使用 LSTM 专门来处理这些金融时间序列数据。即首先使用 CNN 提取特征，卷积核设置同 CNN 模型，得到 32 个 1×5 的张量后拼成 5×32 的二维张量。送入隐藏层状态大小为 16 的一层 LSTM 中得到预测结果。

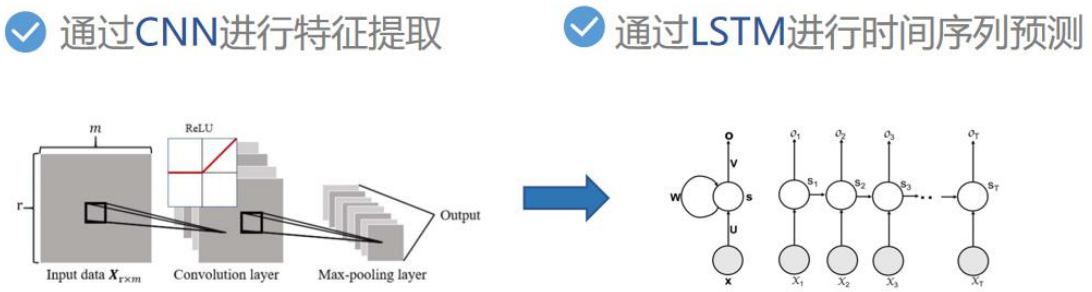

图 15 CNN-LSTM 模型原理示意图

4.2 机器学习模型

4.2.1 支持向量机

作为一种常见的机器学习方法，支持向量机(support vector machines， SVM)在量化领域都得到了广泛的应用。支持向量机的核心思路是在特征空间中构建一个超平面，使得不同类别的样本点距离该超平面的间隔最大。支持向量机采用了将特征空间映射到更高维度的方法来解决原始特征空间中无法做到线性可分的难题。区别于传统的统计方法，支持向量机最小化结构风险，在控制住经验风险的情况下，尽可能缩小置信区间。由于我们的研究目的是预测股价的涨跌，所以考虑二元分类的支持向量机。

设样本集为$(x_i,y_i),x_i \in R^m,y_i \in 0,1,i = 1,2,.....n$考虑设置误差带宽和惩罚函数的情况下，将最大化间隔转化为优化问题：

$$\min \frac{1}{2} w^T w + C \sum_{i=1}^{n} \varepsilon_i$$

$$s.t.\, y_i(w^T \emptyset(x_i) + b) + \varepsilon_i - 1 \geq 0$$

其中，w 与 b 是参数，ε_i是 i 点的松弛变量，C 为惩罚系数，$\emptyset(x_i)$为核函数。求解上述问题即可形成分割超平面。本文将特征输入，将未来股票价格涨跌作为标签进行预测。

4.2.2 人工神经网络

多层感知机（MLP，Multilayer Perceptron）也叫人工神经网络（ANN，Artificial Neural Network）是一种高度数学与计算化的非线性动力学仿生系统。MLP 通过训练模型，不断修正权重从而使输出值不断逼近最优值。训练神经网络的核心即如何从数据中估计所需的权重，本文采用常见的反向传播法来估计模型权重。

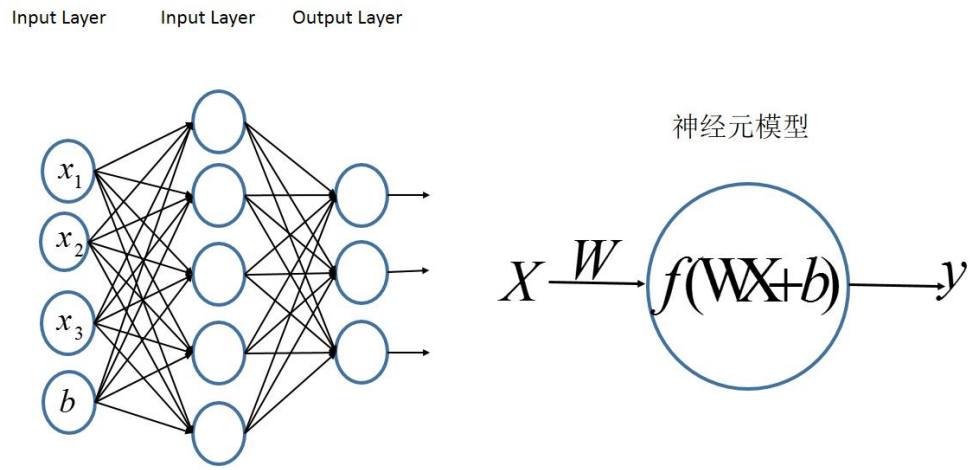

图 16 MLP 模型原理示意图

4.2.3XGBoost 模型

XGBoost 是一个可扩展的树提升系统，该算法引入了正则化项来控制树的复杂度，并在迭代过程中将泰勒展开式应用于目标函数中，这很大程度上加速了模型的优化。此外，它还有支持自定义损失函数、可以处理缺失值、并行计算等优点，越来越受到量化领域学者的关注和欢迎。因此本文结合自身数据，应用 XGBoost 作进一步预测分析。

XGBoost 进行 t 次迭代的目标函数：

$$obj^{'}=\sum_{i=1}^{n} l(y_i,\hat{y}_i)+\sum_{k=1}^{t}\Omega(f_t)=\sum_{i=1}^{n} l\left(y_i,\hat{y}_i^{(t-1)}\right)+f_t(x_i))+\Omega(f_t)+const$$

其中$\sum_{i=1}^{n} l(y_i,\hat{y}_i)$为损失函数，$\Omega(f_t)$为第 t 棵树的复杂度，$const$ 为常数。

目标函数的最终形式为：

$$obj^{(t)}=\sum_{j=1}^{T}[G_j\omega_j+\frac{1}{2}(H_j+\lambda)\bar{\omega}_j^2]+YT$$

4.2.4LightGBM 模型

LightGBM 是近年来提出的一种数据分类模型，其核心为梯度提升树(GBDT)算法。单一的 GBDT 算法只能基于回归树进行计算，因此新树的计算需要借鉴之前所有树的结论和残差，然后基于多个决策树的结果，得到最终的预测输出。传统的 GBDT 算法需要构建一定数量的决策树，划分最优分割点，并对特征值进行排序，这花费了大量的计算时间，降低了计算效率。同时随着目前所有解决问题所含数据量的增加，单一的 GBDT 算法的精度和计算效率无法满足需求，因此需要引入 LightGBM 算法。

相较于传统的 GBDT 算法，LightGBM 算法通过以下几个方面提高了算法的高效性:

(1)将特征值划分为多个区间，在每个区间中选取对应的分割点，这样即提升了计算效率，同时避免了过拟合的现象;

(2)采用 leaf-wise 生长策略，每生长一枚叶子时相比于传统的策略都可以减少损失，同时需要设定额外的参数避免过拟合的现象;

(3)使用特征捆绑方法，高维度数据中存在多个特征值，且特征值之间存在信息冗余问题，特征捆绑法将上述特征值放入稀疏空间中，降低计算复杂性。

4.3 模型评价

均方误差(Mean Squared Eror，简称 MSE)是数据预测中常用的评价指标，MSE 的值越小，误差的离散程度越小，预测效果也就越好。MSE 计算公式如下:

$$MSE = \frac{1}{n}\sum_{t=1}^{n}(Y_t - y_t)^2$$

五、实验及结果分析

5.1 整体流程

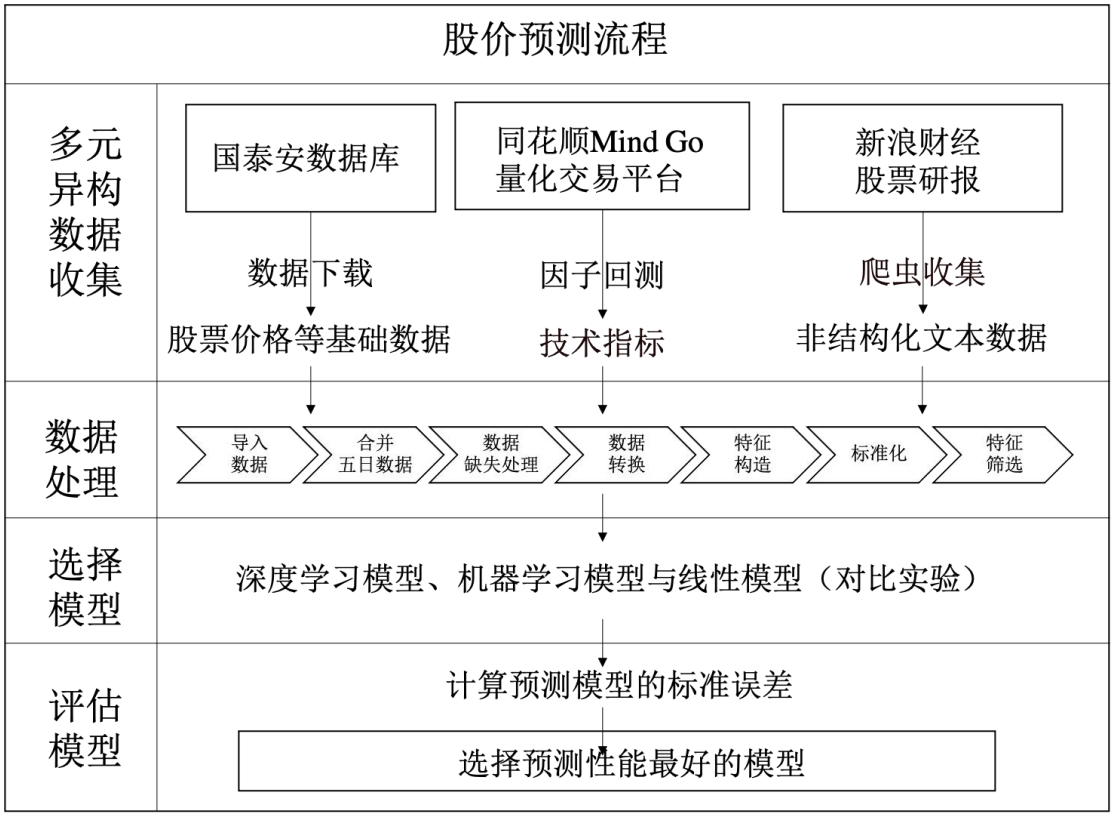

图 17 多源异构数据预测的整体流程

5.2 实验统一设置

本实验中共有 62780 个可用样本。利用 5 个连续交易日的数据预测下一个交易日的数据，使用滑动窗口的方法对数据进行切分，每支股票三年的交易日数据可划为 725 条。按时间排列后，取前百分之九十的数据作为训练集，后百分之十的数据作为测试集，对模型进行训练和测试。训练集样本数为 56502 个，测试集样本数为 6278 个。

在深度学习模型中，设置随机数种子为 1122，最大训练轮数设置为 10，一批数据 20 条。使用早停法防止模型过拟合，若连续 5 轮测试集损失都为下降，就停止训练。损失函数使用 MSELoss，优化器使用 Adam，学习率为 0.0001。

在机器学习模型中，先删除信息熵与情感两类特征再针对不同行业进行训练。

5.3 深度学习模型实现与求解

一般的金融时间序列数据的存储结构中并不存在二维卷积神经所需的有意义的二维空间关系。且使用二维卷积网络处理金融时间序列数据时，存在纵向移动会导致时间上信息的损失的问题。因此本文选择使用一维卷积神经网络模型进行特征提取与预测。设置卷积核的尺寸与输入数据的尺寸同宽，让卷积窗口仅做纵向的滑动能更好适应金融时间序列数据。

本文采用了 CNN 模型、LSTM 模型、CNN-LSTM 混合神经网络分别进行预测对比结果。具体数据处理和特征学习过程如下：

（1）输入层。经过结构化和非结构化数据的处理，本文对构建的量化因子进行筛选后，最终得到 49 个特征。输入的数据为经过（0，1）标准化处理后的样本数据。每一个样本包含 5 个交易日的特征数据，本文使用 5 个交易日预测第 6 个交易日。输入数据的单个样本为 49*5 的矩阵。

（2）卷积层。卷积层的作用是提取一个局部区域的特征，不同的卷积核相当于不同的特征提取器。每批选取 20 条数据，为了使用一维卷积，将数据格式进行转变，将 49 个特征作为宽，天数 5 作为长。滤波器的步长即滑动时的时间间隔设置为 1，使用零填充（zero padding）参数设置为 1。以尺寸为 3 的一维卷积核对输入数据进行卷积处理，卷积核个数为 32，即每一个卷积核对输入空间按照 49*3 的窗口大小横向滑动提取特征，训练 32 个卷积核，共提取出 32 种不同的特征。其中经过一维卷积过后，得到 32 个 1*5 的分布式特征。

对这 32 个张量进行拼接后使用全连接层得到预测结果，即是 CNN 模型的方法。这 32 个张量作为输入数据放入 LSTM 中，使用循环层进行预测，即是 CNN-LSTM 混合神经网络模型的方法。

（3）循环层将卷积层的输出作为该层的输入，即输入数据为特征维度为 32 的数据结构，使用隐藏层状态大小为 16 的一层 LSTM 模型对输入数据进行处理，进行股价的预测。

剔除不受市场影响的小企业的股票进行实验，实验结果如表 3.

表 3 深度学习实验结果

cnn1D（一层卷积无池化层）

MSE	train:0.9 所有 test：0.1 所有	train：0.9 种植业 test：0.1 种植业	train：0.9 化学原料 test：0.1 化学原料	train：0.9 钢铁 test：0.1 钢铁	平均
完整特征 66	3.038416269	0.969350094	6.059466803	0.833293457	2.725131656
去掉情感 63	2.790978395	0.970583831	3.901409451	0.803169398	2.116535269
去掉信息熵 65	2.828525726	0.951207711	5.307856417	0.983280578	2.517717608
去掉情感信息熵 62	2.629605325	1.153336684	2.307915864	1.203217341	1.823518803
去掉文本特征 51	2.765849803	0.995193224	4.615836532	1.176041647	2.388230302
平均	2.810675104	1.007934309	4.438497014	0.999800484	2.314226728

lstm（一层）

MSE	train:0.9 所有 test：0.1 所有	train：0.9 种植业； test：0.1 种植业	train：0.9 化学原料;test：0.1 化学原料	train：0.9 钢铁 test：0.1 钢铁	平均
完整特征 66	2.474699205	1.189232067	3.053862851	0.935942163	1.913434072
去掉情感 63	2.4450411	1.117112853	2.41027684	0.812603747	1.696258635
去掉信息熵 65	2.513052044	1.079240741	2.95255368	1.05214881	1.899248819
去情感信息熵 62	2.47882115	0.968446921	2.279391937	0.663015863	1.597418968
去掉文本特征 51	2.583805058	1.061332279	2.838422515	0.964809417	1.862092317
平均	2.499083712	1.083072972	2.706901565	0.885704	1.793690562

cnn（1 层卷积无池化层）+lstm(1 层)

MSE	train:0.9 所有 test：0.1 所有	train：0.9 种植业； test：0.1 种植业	train：0.9 化学原料;test：0.1 化学原料	train：0.9 钢铁 test：0.1 钢铁	平均
完整特征 66	2.411676088	0.83489926	1.43779116	1.232044652	1.47910279
去掉情感 63	2.536727302	0.883865728	1.674591996	1.439165476	1.633587625
去掉信息熵 65	2.121168359	0.929081971	2.316662377	1.478167093	1.71126995
去情感信息熵 62	2.404599077	0.969142756	1.458767817	1.263899493	1.524102286
去掉文本特征 51	2.37969336	1.013250026	1.448511962	1.244130283	1.521396408
平均	2.370772837	0.926047948	1.667265062	1.3314814	1.573891812

信息熵与情感指标对股价预测的准确度有负面影响。在观察原始数据、特征值分布后本文认为产生原因如下：第一，情感值方面。投资研报与财经新闻、股吧评论等不同。后者发布频率高，时效性短，根据事实的好坏有不同的情感倾向，且十分明显。而前者经常倾向于研究利好趋势，或是使用一些“春秋笔法”。因此

情感值大多为正向，无法客观地反映现实中的情况，也就无法为真正的股价变化提供有效的信息。第二，信息熵方面。从信息熵的值来看，由于投资研报都由专业人员编写，有类似的撰写方式，因此信息熵的值集中在一个很小的区域内，非常相近，可以视为一个无效特征。加入模型中还会为股价的变化带来错误的信息。从信息熵的定义来看，信息熵能够反映一篇文章所包含内容信息程度，侧面反映一篇研报的重要程度；另一方面，信息熵也能够反映一篇文章的冗余程度，过高的信息熵可能会带来相反的效果，即代表研报内容冗余、存在干货不足的可能。因此，仅仅通过一个值，无法判断研报是有效信息多，还是冗余程度大。若想利用这方面的特征，统计学的方法以无法解决，需要结合研报的语义特征来处理。

按行业分类后的预测更为准确。在大多数情况下，用单个行业训练的模型预测此行业内的股票，结果比三个行业混合使用更好。这是因为在同一行业内，各股会受到相同行业波动的影响，因此会产生相似的变化趋势。并且同一行业内的各企业关系密切，比如产品互为替代品的竞争关系，因此各股之间的变化可能是相互牵连的。如果用一个行业的股票训练专门预测此行业股价的模型，模型会学到内在关系，从而产生更好的预测结果。

文本特征对结果的提升效果不显著。加与不加文本特征，对效果没用特别显著的影响。分析原因如下：第一，新浪财经网站的研报数据具有一定范围的局限性。此网站无法涵盖所有股票研报报告，并且研究员在撰写研报时根据自己的主观想法，无法保证绝对的客观和全面。并且只用一种文本数据（投资研报）作为文本信息的来源稍显单薄。第二，数据处理的问题。文本特征存在大量缺失值，用 0 填充的方法未考虑研报信息的时效性、相互作用的关系等因素，因此无法真正地发挥研报数据对股价预测的促进作用。文本量化只有 11 个浅层特征与 3 个情感特征、1 个信息熵特征，未把研报中所有有价值的信息挖掘出来。第三，模型问题。模型的结构不适合处理大量缺失的数据，可以从模型内部结构做出改进。

CNN 的效果优于 LSTM，而本文将二者结合的模型具有最好的效果。CNN 的卷积操作可以提取数据中的有效特征，为其加权，并减小特征数。LSTM 适合处理时序数据，通过 Cell 状态保留数据之间的时序关系。因此通过 CNN 提取特征并简化，同时保留其中的时序关系，送入 LSTM 中进行最后的预测，达到了最好的效果。

5.4 机器学习模型实现与求解

整体来说，XGBoost 模型、LightGBM 模型与 MLP 模型直接运用默认参数进行训练发现其训练效果较好，SME 均小于 1。图 18-21 展示各模型预测效果与真实情况对比图。其中，MLP 神经网络的性能最好，其次是 LightGBM，最差的是 XGBoost。本文分析原因认为，首先 LightGBM 本身即作为 XGBoost 的升级加强版本，且无论是训练速度还是内存消耗都明显具有优势。LightGBM 在 XGBoost 基础上采用 Histogram、GOSS、EFB 算法对 XGBoost 训练性能进行了优化。而对于神经网络而言，它具有高度的并行性。神经网络是由许多相同的简单处理单元并联组合而成，虽然每个单元的功能简单，但大量简单单元的并行活动，使其对信息的处理能力与效果惊人。

由于 SVM 模型使用默认参数得到的 MSE 较大，故对模型进行参数调整。使用贪心算法，并且选择‘neg_mean_squared_error’作为评价标准，对惩罚参数 C 以及核函数 kernel 进行调整。图 22 为绘制参数选择迭代过程中的 MSE 结果情况图。经过参数调整后，模型预测的 MSE 有所下降，但效果仍然没有达到一个较好的效果。表 4 展示各机器学习模型 MSE 结果。

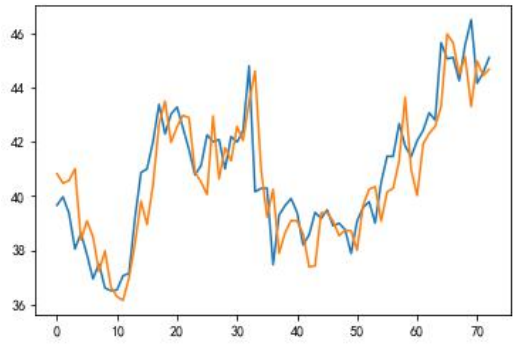

图 18XGBoost 预测效果与真实情况对比图

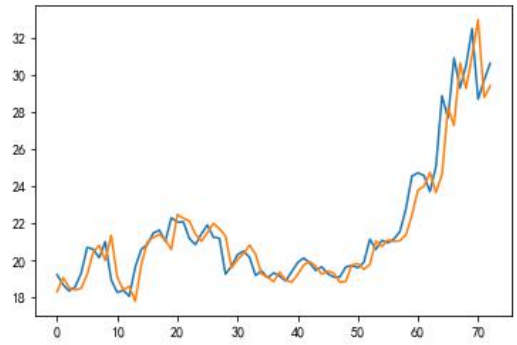

图 19LightGBM 预测效果与真实情况对比图

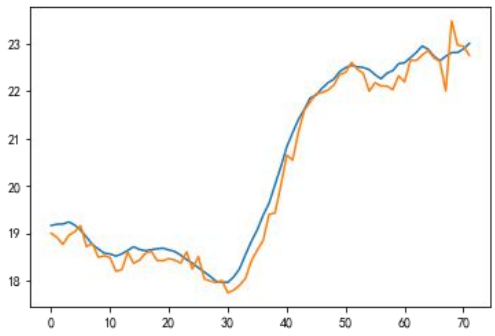

图 20 MLP 预测效果与真实情况对比图

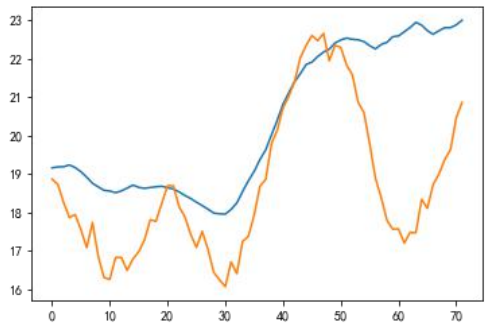

图 21 SVM 预测效果与真实情况对比图

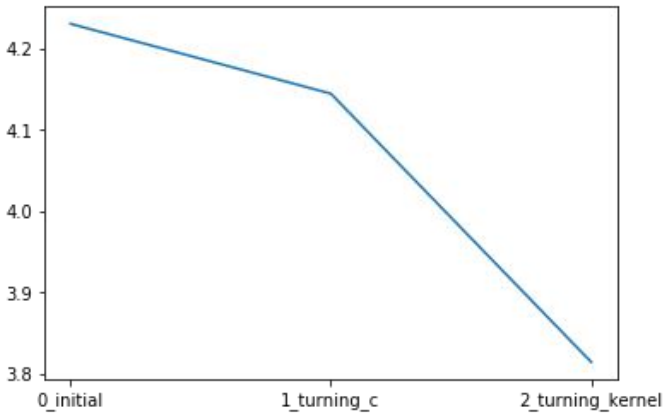

图 22 绘制参数选择迭代过程中的 MSE 结果情况图

表 4 机器学习实验结果

全部公司					
MSE	train:0.9 所有 test：0.1 所有	train：0.9 种植业 test：0.1 种植业	train：0.9 化学原料 test：0.1 化学原料	train：0.9 钢铁 test：0.1 钢铁	平均
XGBoost					
完整特征 66	0.165335643	0.183354252	0.208515091	3.864789261	1.105498562
去掉情感 63	0.180776863	0.183354252	0.218438783	3.99182928	1.143599795
去掉信息熵 65	0.173321807				0.173321807
去掉情感和信息熵 62					0
平均	0.173144771	0.183354252	0.213476937	3.928309271	0.605605041
LightGBM					
MSE	train:0.9 所有 test：0.1 所有	train：0.9 种植业；test：0.1 种植业	train：0.9 化学原料；test：0.1 化学原料	train：0.9 钢铁；test：0.1 钢铁	平均
完整特征 66	0.174598612	0.155064005	0.278653254	2.112788532	0.680276101
去掉情感 63	0.174015963	0.155064005	0.262589939	2.096834433	0.672126085
去掉信息熵 65	0.173050151				0.173050151
去掉情感和信息熵 62					0
平均	0.173888242	0.155064005	0.270621596	2.104811483	0.381363084
SVM					
MSE	train:0.9 所有 test：0.1 所有	train：0.9 种植业；test：0.1 种植业	train：0.9 化学原料；test：0.1 化学原料	train：0.9 钢铁；test：0.1 钢铁	平均

完整特征 66	3.814664079	1.3822805	2.501114775	5.189419404	3.22186969
去掉情感 63	3.866661015	1.3822805	2.52840686	4.956799889	3.183537066
去掉信息熵 65	3.848152811	1.382408705	2.433287002	9.999701846	4.415887591
去掉情感和信息熵 62	3.892948147	1.382408705	2.767960739	9.820724883	4.466010618
平均	3.855606513	1.382344603	2.557692344	7.491661506	3.821826241

神经网络					
MSE	train:0.9 所有 test：0.1 所有	train：0.9 种植业；test：0.1 种植业	train：0.9 化学原料；test：0.1 化学原料	train：0.9 钢铁；test：0.1 钢铁	平均
完整特征 66	0.058487454	0.350394052	0.613576909	1.908093952	0.732638092
去掉情感 63	0.22317757	0.480547788	0.359839923	0.716194903	0.444940046
去掉信息熵 65	0.335575102	0.342539912	1.171007271	0.713097919	0.640555051
去掉情感和信息熵 62	0.055448741	0.291602501	0.383134935	0.62929647	0.339870662
平均	0.168172217	0.366271063	0.631889759	0.991670811	0.539500963

不同于深度学习模型，在机器学习模型中综合行业训练结果优于各行业分开训练结果。这可能因为综合行业为训练提供了更加丰富的数据样本，能使模型更加良好的学习股价和特征之间的回归关系，具有更好的泛化能力，所以提升了模型表现。

情感和信息熵特征对种植业影响较小但对其他行业结果有明显影响。与深度学习结果类似，针对 LightGBM 模型，数据集在去除情感和信息熵两类特征后表现更佳。但是对于 XGBoost 和 MLP 模型，去除情感和信息熵两类特征后表现变差，但是变化较小。而对于种植业的股票而言，删除情感或信息熵特征后，模型效果并没有大的改变，我们猜测是因为种植业由于交易量较少，针对该行业的研报分析都较为保守，情感和信息熵都集中于一个较小的区间，没有强有力的信息对股价造成影响，相当于两类无效特征。

SVM 的使用具有局限性。本实验使用到的机器学习模型都具有较好的结果，但 SVM 在数据量较大的股票预测中表现很差。SVM 模型之所以能成为目前最常用，效果最好的分类器之一，在于其优秀的泛化能力；由于其本身的优化目标是结构风险最小化，所以可以降低了对数据规模和数据分布的要求。因此 SVM 在小样本训练集上可以取得更好的效果。而对于本实验数据，SVM 的效果较差。

5.5 线性模型实现与求解

运用线性回归模型，对机器学习与深度学习模型效果进行对比。本文分别采用岭回归，弹性网络回归及偏最小二乘回归进行回归，发现线性模型的结果总体较差，MSE 在 15 以上。这是因为实际情况应是非线性关系，不适合运用线性模型进行估计。线性模型回归结果如表 5 所示。

图 5 线性模型实验结果

线性模型	MSE
线性模型	6.37E+20
岭回归	18.18098008
弹性网络回归	56.86676412
偏最小二乘回归	76.37840339

在本实验中，机器学习模型的效果优于其他线性模型及深度学习模型，其中 MLP 神经网络的效果最佳。在实验使用到的模型结果中，机器学习模型的平均结果为 0.171735077，深度学习模型结果平均值为 2.370772837，线性模型的平均结果为 1.59E+20。分析原因可能有以下几点：第一，数据量较少。对于深度学习而言，使用到的数据需要较多的特征以及较大的数据量。由于本实验使用到的数据对比需要的数据量较少，所以深度学习的效果不佳。第二，受缺失值的影响较大。由于本实验的数据是通过爬虫获取，故可能存在很多缺失值，对模型的评价效果会有很大影响。第三，模型结构设置不合理。训练深度学习模型需要耗费较多的时间和较大的计算机资源，故在模型结构设置尝试过程中无法尝试到最优的结构，导致结果有并非深度学习可以达到的最好效果。

情感特征和信息熵对结果的影响与模型有很强的关系。对于深度学习和 LightGBM 等模型，数据集在去除情感和信息熵两类特征后表现更佳，说明这类模型对于股票数据中的情感特征和信息熵的拟合效果较差。但是对于 XGBoost 和 MLP 等模型，去除情感和信息熵两类特征后表现会稍微变差，故此类模型应使用情感特征和信息熵进行进一步分析。

六、结论

本文主要基于多源异构数据的分析方法，对国内股指数据、研报数据与新闻数据进行研究，以 A 股钢铁行业、化学原料行业以及种植业为例，用五天的历史数据来预测第六天的股价或者涨跌情况，实现了对未来股价的初步预测工作。与线性模型进行对比，本文选择的模型预测结果准确度均相对较高。在深度学习模型中，分别使用 LSTM、CNN 与 CNN-LSTM 模型进行求解，最终最佳效果为 CNN-LSTM 模型，MSE 评价为 0.83489926。在机器学习模型中，分别使用 MLP 神经网络、LightGBM、XGBoost 与 SVM 模型进行预测，其中效果最好的是 MLP 神经网络，MSE 评价为 0.539500963。

对模型求解的后验分析发现，研报处理出来的情感与信息熵对模型评估有负面作用，考虑是因为缺失值过多，以及特征构造不够充分；文本特征对模型提升效果不明显，因为所收集的新闻与研报数据过少，特征稀疏，不能够给模型带来较明显的作用。其它方面，整体建模会对微型上市企业预测带来过大影响，因为行业特征会影响其预测趋势，但微型企业的股价几乎不波动，后续考虑剔除微型企业，整体建模，效果有较大提升。

最终我们实现了基于多源异构数据的股价预测，对收集数据、处理数据、模型求解、模型评估以及修正等理论方法有进一步的学习理解，创新的实现了 CNN 与 LSTM 模型的融合，模型能够一定程度上进行选股与模拟交易策略，在量化交易方面做出了进一步贡献。

参考文献

[1]季子峥，沈婷婷，张孝.利用社交媒体情感分析的短期股价趋势预测方法[J].北京理工大学学报，2020，40(01)：83-89.
[1]赵丽丽，赵茜倩，杨娟，王铁军，李庆.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(07)：70-75+80.
[3]Nader Mahmoudi, Paul Docherty, Pablo Moscato Deep neural networks understand investors better [J] Decision Support Systems，2018，112
[4]ChauF, Deesomsak R, Koutmos D. Does investors entiment really matter? [J] International Review of Financial Analysis, 2016, 48：221-232.
[5]饶东宁，邓福栋，蒋志华.基于多信息源的股价趋势预测[J].计算机科学，2017，

44(10)：193-202.
[6]LAM M. Neural network techniques for financial performanceprediction:integrating fundamental and technical analysis[J].Decision Support Systems，2004，37(4):567-581.
[7]LI X，HUANG X，DENG X，et al.Enhancing quantitative intra- day stock return prediction by integrating both market news and stock prices information[J].Neurocomputing，2014，142(1):228-238.
[8]DUAN J， ZENG J.Forecasting stock return using multiple information sources based on rules extraction[C]‖12th Interna- tional Conference on Fuzy Systems and Knowledge Discovery (FSKD 15).Piscataway，New Jersey:IEEE，2015:1183-1188
[9]DUAN J，LIN H，ZENG J. Posterior probability model for stock return prediction based on analyst's recommendation behavior[J].Knowledge-Based Systems，2013，50:151-158.
[10]NEWMAN M R，GAMBLE G O，CHIN W W，et al. An Investigation of the Impact Publicly Available Accounting Data，Other Publicly Available Information and Management Guidance on Analysts' Forecasts[M]‖ New Perspectives in Partial Least Squares and Related Methods. New York: Springer，2013:315-339
[11]LEE M C. Using support vector machine with a hybrid feature selection method to the stock trend prediction[J].Expert Systems with Applications， 2009， 36(8):10896-10904.
[12]BOLLEN J， MAO H， ZENG X. Twiter mood predicts the stock market[J].Computer Science，2010，2(1):1-8.
[13] 徐琳. 网络舆情对股价波动影响的实证研究[D].西南财经大学,2013.
[14]TSAI C， LIN Y， YEN D C， et al. Predicting stock returns by clasifier ensembles[J].Applied Soft Computing，2011，11(2):2452-2459.
[15]LAM M. Neural network techniques for financial performance prediction: integrating fundamental and technical analysis[J].Decision Support Systems，2004，37(4):567-581.
[16]PATEL J，SHAH S，THAKKAR P，et al. Predicting stock and stock price index movement using Trend Deterministic Data Preparation and machine learning techniques[J].Expert Systems with Applications，2015，42(1):259-268.
[17]林培光,周佳倩,温玉莲.SCONV:一种基于情感分析的金融市场趋势预测方法[J].计算机研究与发展,2020,57(08):1769-1778.
[18]赵丽丽,赵茜倩,杨娟,王铁军,李庆.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(07):70-75+80.
[19]张梦吉,杜婉钰,郑楠.引入新闻短文本的个股走势预测模型[J].数据分析与知识发现,2019,3(05):11-18.
[20]王燕,郭元凯.改进的 XGBoost 模型在股票预测中的应用[J].计算机工程与应用,2019,55(20):202-207.
[21]王晓红,王梦瑶,郝婷.改进的时间相关序列股票价格混合预测模型研究[J].科技促进发展,2020,16(06):672-678.
[22] PATEL J, SHAH S, THAKKAR P, et al. Predicting Stock and Stock Price Index Movement Using Trend Deterministic Data Preparation and Machine Learning Techniques[J]. Expert Systems with Applications, 2015, 42(1):259-268.

[23]曹晓，孙红兵 ．基于灰色 GARCH 模型和 BP 神经网络的股票价格预测[J].软件,2017，38(11):126-131.
[24]ANGGRAENI W , ANDRI K B , SUMARYANTO, et al. The Performance of ARIMAX Model and Vector Autoregressive (VAR) Model in Forecasting Strategic Commodity Price in Indonesia[J]. Procedia Computer Science, 2017, 124:189-196.
[25] NEZHAD S M T, NAZARI M, GHARAVOL E A. A Novel DoS and DDoS Attacks Detection Algorithm Using ARIMA Time Series Model and Chaotic System in Computer Networks[J]. IEEE Communications Letters, 2016, 20(4):700-703.
[26] WANG S X, ZHANG N, WU L, et al. Wind Speed Forecasting Based on the Hybrid Ensemble Empirical Mode Decomposition and GA-BP Neural Network Method[J]. Renewable Energy, 2016, 94: 629-636.
[27] YOU J, WANG J, FANG S, et al. An Optimized Real-Time Crash Prediction Model on Freeway with Over-Sampling Techniques Based on Support Vector Machine[J]. Journal of Intelligent & Fuzzy Systems, 2017, 33(1):555-562.
[28]张晨希,张燕平,张迎春,陈洁,万忠.基于支持向量机的股票预测[J].计算机技术与发展,2006(06):35-37.
[29]Lapedes A, Farber R. Nonlinear signal processing using neura I networks; Prediction and system modelling[R]. Los Alamos: Los Alamos National Laboratory, 1987.
[30]孙存浩,胡兵,邹雨轩.指数趋势预测的 BP-LSTM 模型[J].四川大学学报(自然科学版),2020,57(01):27-31.
[31]赵红蕊,薛雷.基于 LSTM-CNN-CBAM 模型的股票预测研究[J].计算机工程与应用,2021,57(03):203-207.
[32]文宇.基于 CNN-LSTM 网络分析金融二级市场数据[J].电子设计工程，2018，26:75.
[33]胡婧,叶建木.基于微博信息的股票交易预测研究[J].财政监督,2017(05):108-111.
[34]孔翔宇,毕秀春,张曙光.财经新闻与股市预测——基于数据挖掘技术的实证分析[J].数理统计与管理,2016,35(02):215-224.
[35]王亚红,程希明.基于财务指标的股价预测模型及实证研究[J].区域金融研究,2018(09):35-38.
[36]李斌,邵新月,李玥阳.机器学习驱动的基本面量化投资研究[J].中国工业经济,2019(08):61-79.
[37]李斌,林彦,唐闻轩.ML-TEA:一套基于机器学习和技术分析的量化投资算法[J].系统工程理论与实践,2017,37(05):1089-1100.
[38]Chen, Y. W., Chen, K., Yuan, S. Y., & Kuo, S. Y. (2016). Moving object counting using a tripwire in H. 265/HEVC bitstreams for video surveillance. Ieee Access, 4, 2529-2541.

附录一：数据可视化

1 结构化数据处理

1.1 数据描述性统计分析

通过三个行业的股市 K 线数据，能够初步得知行业在近三年内的股价波动情况以及行业整体变化，能够对后续工作建立初步探索与认识。对于基础 K 线数据，三个行业的基本数据描述性统计分析结果如下：

Table 1 钢铁行业基本数据描述性统计结果

	open	high	low	close	preclose	volume	amount	turn	pctChg
mean	6.240355	6.344876	6.146907	6.246490	6.245810	3.463070e+07	1.776242e+08	1.132182	-0.000138
std	4.514298	4.615145	4.431133	4.527743	4.516018	6.019549e+07	3.083422e+08	1.261160	2.276565
min	1.040000	1.070000	1.040000	1.050000	1.050000	0.000000e+00	0.000000e+00	0.001956	-10.112360
25%	3.310000	3.350000	3.270000	3.310000	3.310000	6.401910e+06	3.285440e+07	0.370005	-1.094681
50%	4.840000	4.910000	4.770000	4.840000	4.840000	1.839977e+07	7.869681e+07	0.706868	0.000000
75%	7.700000	7.870000	7.560000	7.720000	7.720000	4.095853e+07	1.974350e+08	1.408768	1.023113
max	50.500000	54.220000	49.530000	54.220000	50.550000	1.811060e+09	7.482788e+09	15.098500	10.169500

Table 2 化学原料基本数据描述性统计结果

	open	high	low	close	preclose	volume	amount	turn	pctChg
mean	9.042592	9.227777	8.875259	9.050525	9.051959	1.711293e+07	1.149292e+08	2.008151	0.002091
std	7.580783	7.786596	7.402381	7.592042	7.589264	4.001627e+07	2.915607e+08	2.748831	2.706357
min	1.310000	1.320000	1.300000	1.310000	1.310000	0.000000e+00	0.000000e+00	0.007800	-13.403500
25%	4.530000	4.600000	4.450000	4.540000	4.540000	3.187226e+06	2.196533e+07	0.610623	-1.321596
50%	6.900000	7.000000	6.790000	6.900000	6.900000	7.016596e+06	4.995892e+07	1.113090	0.000000
75%	10.690000	10.930000	10.510000	10.730000	10.740000	1.682434e+07	1.156659e+08	2.238024	1.245403
max	58.790000	60.680000	57.320000	58.440000	58.440000	1.126629e+09	1.202904e+10	52.687932	19.988800

Table 3 种植业基本数据描述性统计结果

	open	high	low	close	preclose	volume	amount	turn	pctChg
mean	8.973875	9.193097	8.806625	9.006014	9.001554	2.271825e+07	2.056679e+08	3.744096	0.082051
std	4.952246	5.107602	4.842262	4.979712	4.973248	2.932984e+07	3.071057e+08	4.905663	3.148524
min	2.030000	2.060000	1.930000	2.060000	2.060000	0.000000e+00	0.000000e+00	0.037871	-11.692300
25%	5.010000	5.110000	4.910000	5.020000	5.012500	5.410970e+06	3.339385e+07	0.893700	-1.472951
50%	8.120000	8.290000	7.960000	8.140000	8.135000	1.262963e+07	9.549523e+07	1.951132	0.000000
75%	11.490000	11.770000	11.247500	11.510000	11.510000	2.820829e+07	2.474215e+08	4.508800	1.377892
max	36.510000	36.510000	36.510000	36.510000	36.510000	3.485447e+08	4.234999e+09	47.751965	15.197400

此外，从其它数据部分可以得知，A 股共计钢铁行业个股 34 支，化学原料个股 33 支，种植业个股 19 支。

根据结果可以初步看出，化学原料的平均股价与最大股价均最高，但标准差较大，相对而言钢铁行业平均股价较低，但总体差异相对较小。三个行业均有成交量为 0 的个股，存在影响模型的可能，在后续处理中应当删除或对其进行特殊处理。在成交量与成交总额规模上来看，三个行业比较接近，钢铁行业的成交规模相对较大，种植业的成交规模差距相对较大。换手率来看，种植业换手率相对较高，钢铁行业则相对较低。

总的来说，三个行业整体市场表现相近，交易相对活跃，具有较好的研究价值。

1.2 结构化数据可视化

对于三个行业的相对占比，绘制饼状图如下所示：

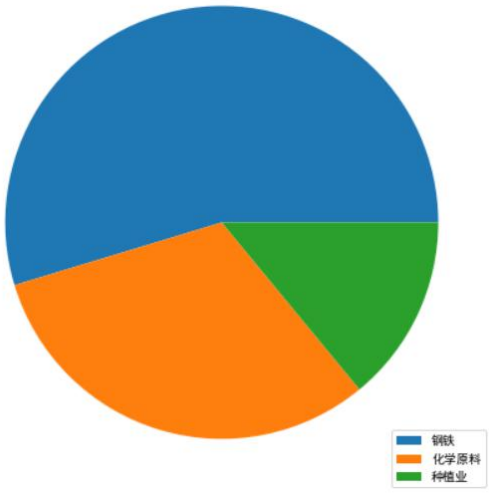

Figure 1 三行业成交总额相对占比饼状图

其中最上面为钢铁，左下角为化学原料，右下角为种植业。

通过饼状图可以看出，钢铁行业的成交总额占比最大，化学原料其次，种植业最少，这与各个行业的相对股票个数有关。去除股票个数影响后，相对占比如下所示：

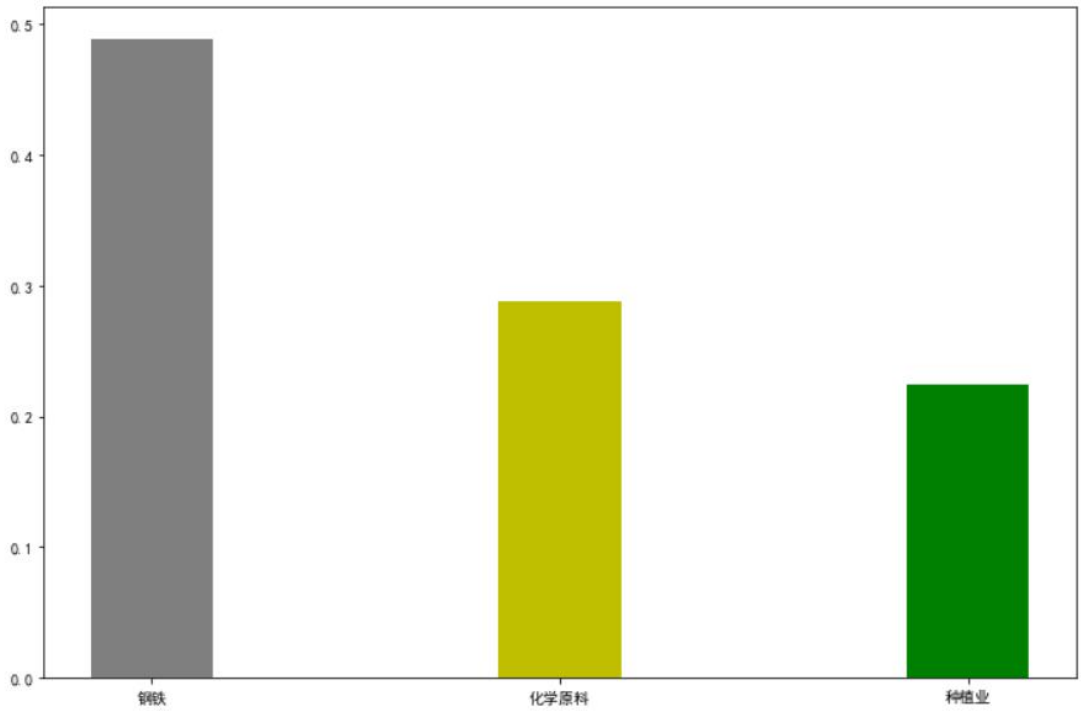

Figure 2 三行业成交总额相对占比饼状图

可以看出，钢铁行业整体平均占比也最大，化学原料其次，种植业最少。在研究过程中，可以分别认为钢铁行业代表规模相对较大的行业股市模型，化学原料行业代表规模相对中等的行业股市模型，种植业代表规模相对较小的行业股市模型。

研究三个行业的日 K 线数据，首先对整体进行研究，绘制 K 线图如下所示：

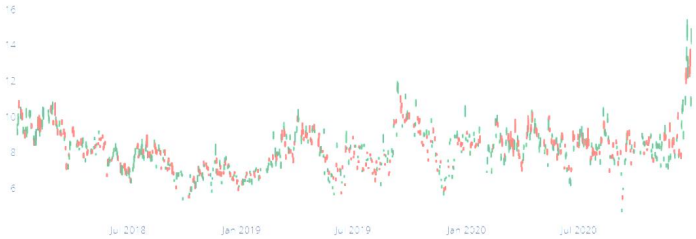

Figure 3 钢铁行业 K 线图

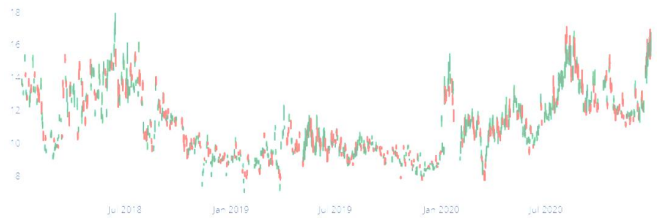

Figure 4 化学原料 K 线图

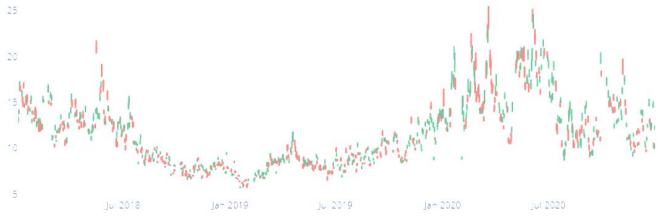

Figure 5 种植业 K 线图

为了对比整体行业情况，特别给出 A 股整体趋势图，如下所示

Figure 6 A 股整体趋势图

初步可以看出，行业股票走势与总体走势有一定部分的相似性，这初步验证了我们先验认为总体指标能够影响行业股票走势的假设。

在股市研究中，股市技术指标为研究员提供了良好的依据，研究员可以通过多种类型的技术指标，判断股票未来可能的趋势，做出买入或卖出的判断。对于所采用的股市技术指标，绘制其在三年内的趋势如下所示：

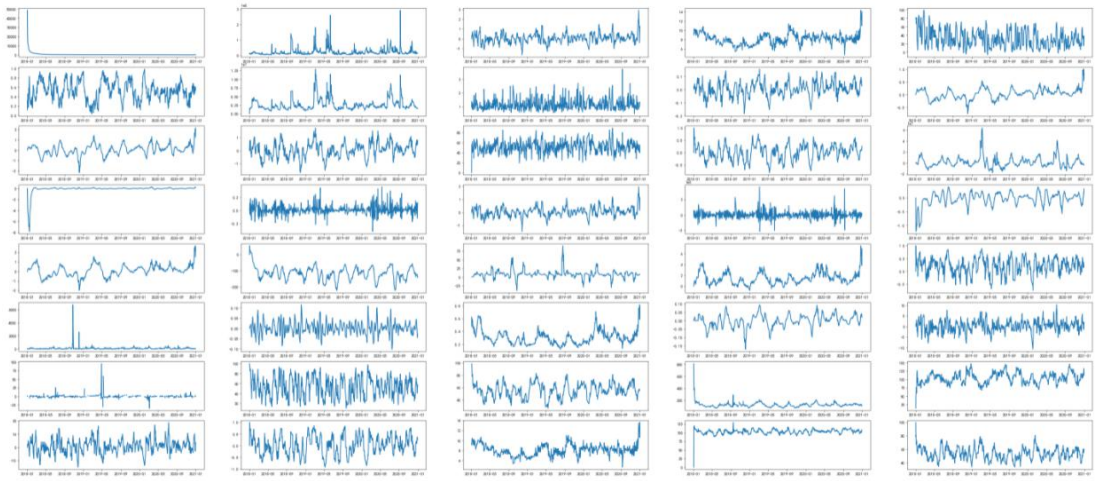

Figure 7 钢铁行业股市技术指标

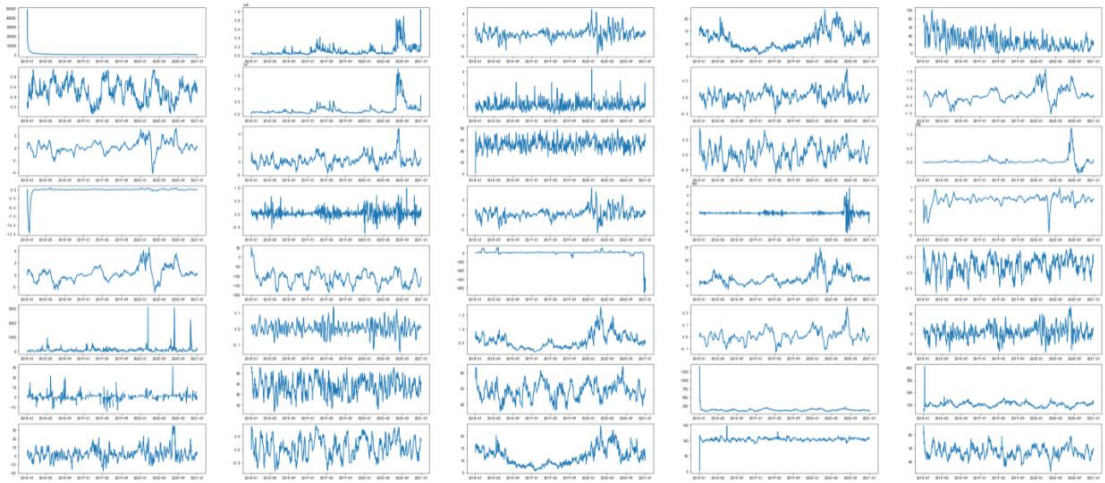

Figure 8 化学行业股市技术指标

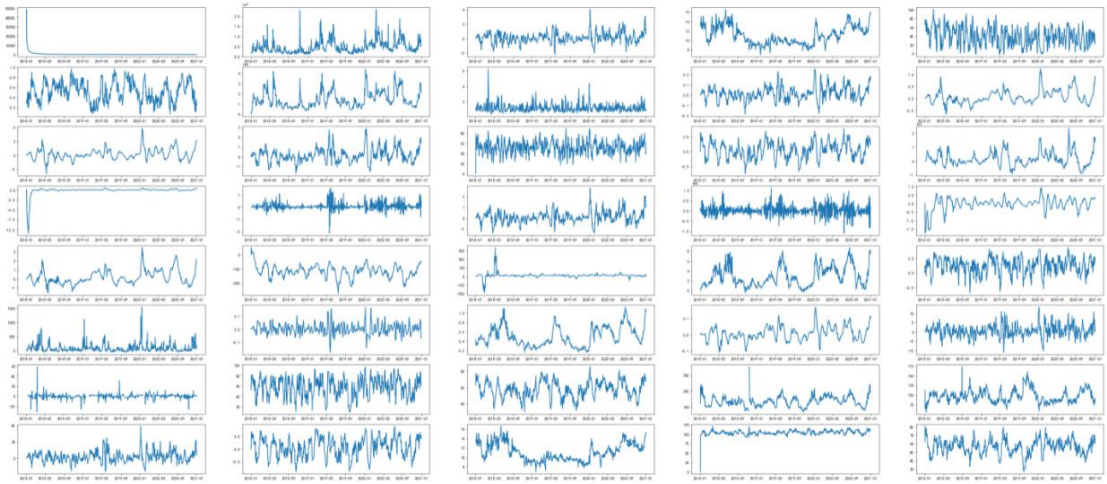

Figure 9 种植业股市技术指标

对于完整的股市技术指标，在因子构造部分会进行详细描述。大体可以看出，各个指标趋势均呈现波动态势，对比而言能够捕捉股市中的波动信息。对

于这些因子，需要进行进一步的因子有效性检测，对检测结果进行分析，筛选出较好结果的因子进行后续分析。

2. 非结构化数据处理

2.1 非结构化数据描述性统计分析

按价值分析的思路，对非结构化数据进行因子构造如下图所示

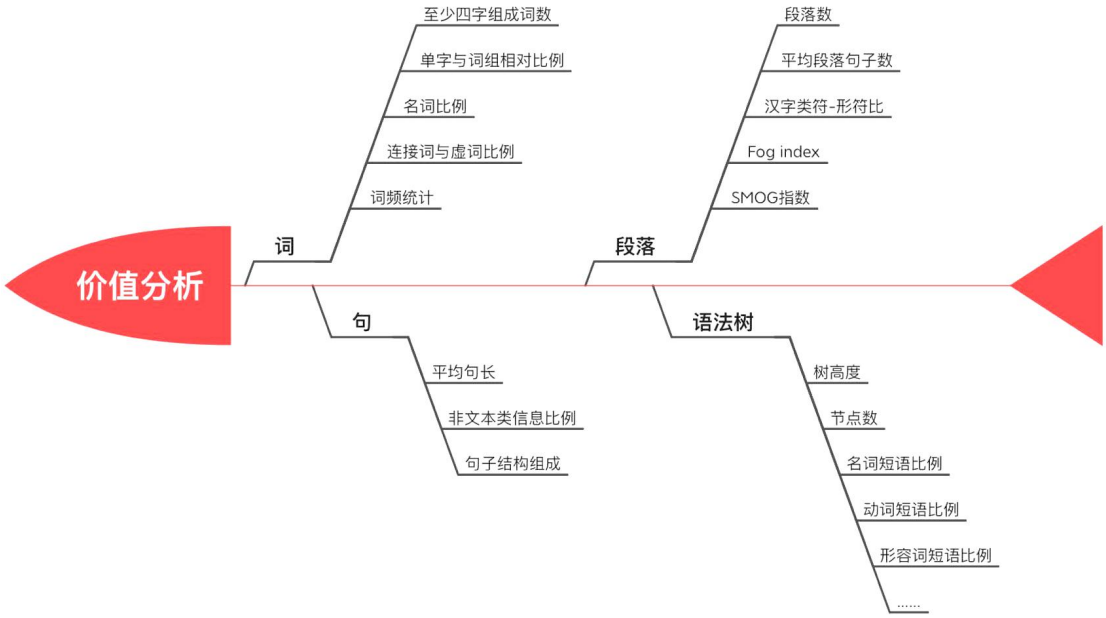

Figure 10 非结构化数据构造因子鱼骨图

最终得到的因子，通过因子有效性检测后，保留 16 个有效因子，其中对于因子名称进行给定并解释如下：

Table 4 非结构化数据构造因子名称释义表

非结构化数据因子名称	因子释义
avg_sentence	平均句长
avg_stroke	平均笔画数
four_word	至少四字组成词数
function_ratio	连接词数与虚词数
n_ratio	名词比例
nine_stroke	至少九笔字数

no_text	非文本信息比例
sen_per_para	平均段落句子数
sum_para	段落数
word_max	每句词最大频率
word_mean	每句词平均频率
word_phrase	单字与词组比例
word_var	每句词频率方差
quality_val	信息熵
pos_num	积极情感词数
neg_num	消极情感词数

对于上述因子，进行描述性统计分析结果如下表所示

Table 5 非结构化数据描述性统计分析

	mean	std	min	25%	50%	75%	max
avg_sentence	86.514857	151.633947	0.0	0.0	0.0	109.598684	1173.116909
avg_stroke	13.322164	23.643381	0.0	0.0	0.0	15.220779	164.857515
four_word	0.048907	0.086600	0.0	0.0	0.0	0.060025	0.566168
function_ratio	0.121677	0.209245	0.0	0.0	0.0	0.161241	1.499095
n_ratio	0.420929	0.735430	0.0	0.0	0.0	0.517241	5.194628
nine_stroke	0.422189	0.747420	0.0	0.0	0.0	0.506550	5.059347
no_text	0.463204	0.862659	0.0	0.0	0.0	0.535610	6.425982
sen_per_para	6.167862	11.397189	0.0	0.0	0.0	7.812500	81.866392
sum_para	17.269497	30.277005	0.0	0.0	0.0	23.000000	237.000000
word_max	23.516884	40.971631	0.0	0.0	0.0	32.000000	303.000000
word_mean	2.994463	5.301544	0.0	0.0	0.0	3.743323	36.900841
word_phrase	0.456381	0.844753	0.0	0.0	0.0	0.517165	6.547497
word_var	5.144481	9.332853	0.0	0.0	0.0	6.670065	71.370224
quality_val	0.917431	2.372050	0.0	0.0	0.0	0.000000	14.804044
pos_num	3.664431	10.229447	0.0	0.0	0.0	0.000000	97.000000
neg_num	1.418429	4.002717	0.0	0.0	0.0	0.000000	34.000000

非结构化数据的最大特点之一就是其稀疏性。从表格数据可以看出，几

乎所有的非结构化数据都存在超过 50%以上的缺失情况，这是行业研究报告与个股研究报告的稀疏性所导致的。在这种情况下，数据处理提供三种解决方案：

（1）不进行处理，通过模型自我识别特征，保留数据真实性；

（2）对数据进行高斯平滑处理，降低其稀疏性，保证数据有效性；

（3）对数据进行基于信息熵权重的记忆累加，即对于某个时刻的研报因子数据，在第二天不会归零，而是继续保留，但会衰减。衰减速率基于信息熵大小确定——信息熵越大，代表非结构化数据越有价值，衰减速率也越小，保证数据记忆性。

经过仔细对比讨论，结合实际数据集，我们选择第一种方案。对于本次数据，我们考虑通过给予模型多源异构数据进行量化交易分析，对于非结构化数据部分，起到一个信息增强作用。对于某个时间点，非结构化数据对个股的股价变化趋势有增强作用。另外，所使用的深度学习 LSTM 模型与 CNN 模型能够识别稀疏信息的特征，已经保证了数据的有效性；同时模型评估过程需要基于真实的数据来推导结论，故在该数据处理中，考虑方案一。另外两种方案也存在一定合理性，保留进一步探讨空间。

2.2 非结构化数据可视化

对于非结构化数据所得到的因子，进行可视化处理，通过图像数据进行进一步探索性分析。

字平均笔画数而言，通过论文研究表明，文本内平均笔画数越大，一般代表着该文本越难懂，并且笔画数一定程度上服从Γ分布。其直方图与核密度图如下所示

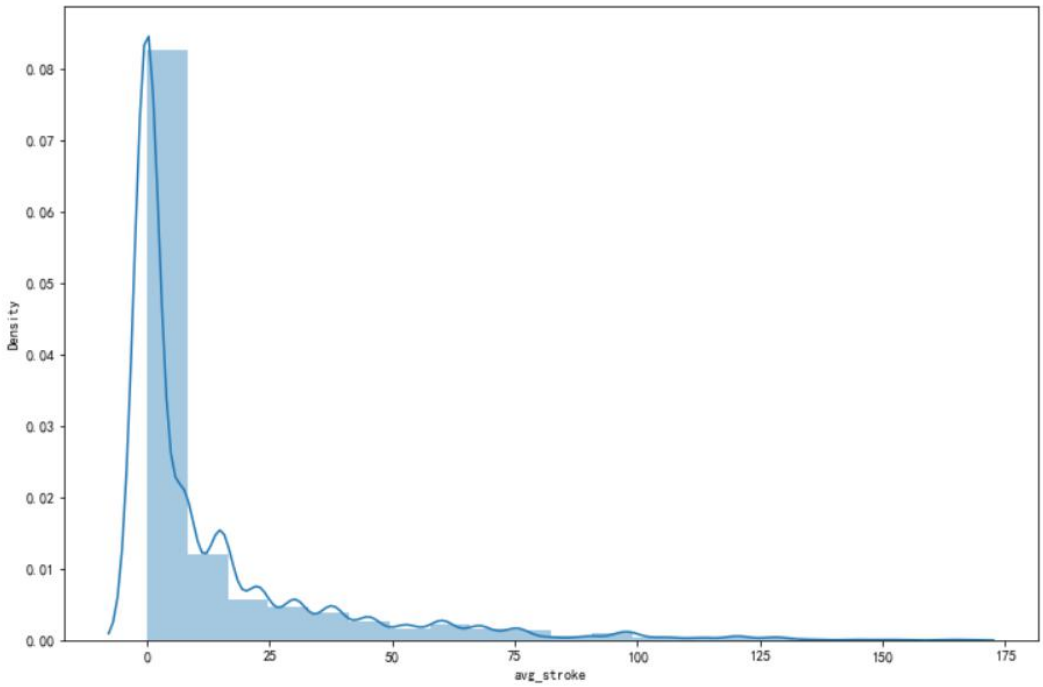

Figure 11 字平均笔画数直方图与核密度图

从图中我们可以看出，除去缺失部分数据，字平均笔画数呈现明显右拖尾现象，对比一般情形的分布，我们可以发现，相较于常规文章而言，股市行业研报相对更加复杂，阅读难度较高，解析难度也相对较高，需要进行更加细致、准确的处理分析。

对于其词频统计，考虑其去除停用词后的均值。在均值分布中，如果平均词频越大，说明词重复次数越多，对于一篇文章，过多的重复与过少的重复都代表该行业研报的质量不高。对于三个行业的研报内容，分析其词频并计算词频均值，绘制直方图与核密度图如下图所示：

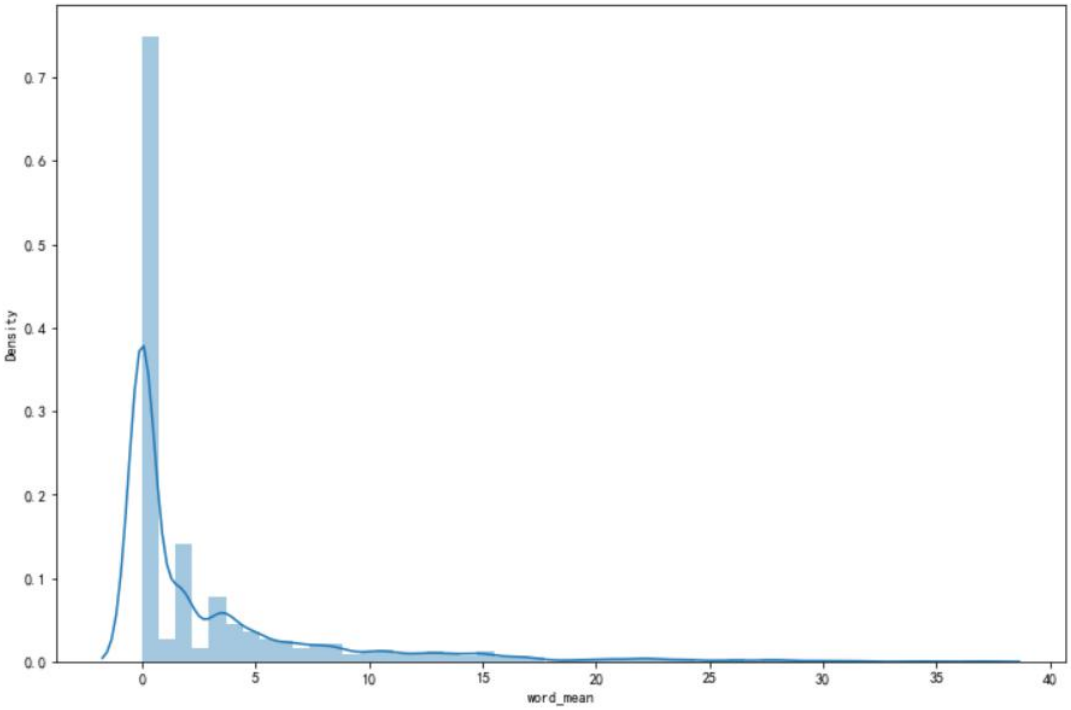

Figure 12 词频均值直方图与核密度图

此外，情感与信息熵是我们所构造的最为重要的两个非结构化数据因子。信息熵能够反映一篇文章所包含内容信息程度，侧面反映一篇研报的重要程度；另一方面，信息熵也能够反映一篇研报内容的冗余程度，过高的信息熵可能会带来相反的效果，即代表研报内容冗余、存在干货不足的可能，在生活中。经过计算，信息熵分布的直方图与核密度图如下所示：

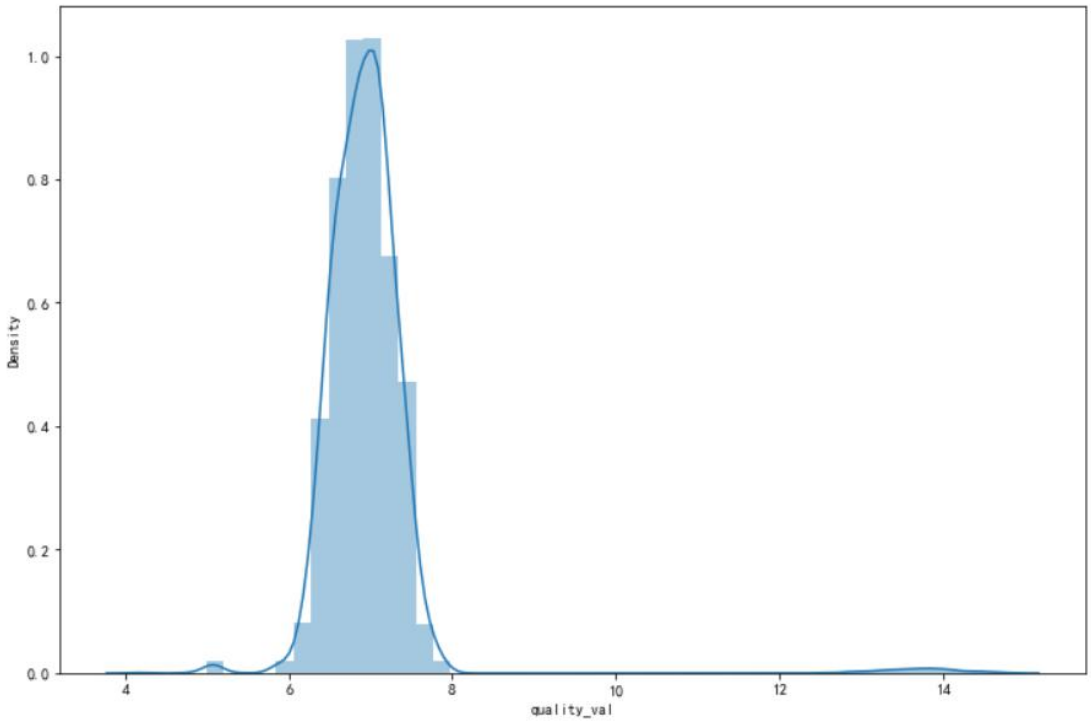

Figure 13 信息熵直方图与核密度图

而情感分析部分，通常分析文本情感是基于词典对照或者词袋模型-神经网络模型进行学习，此处我们选择金融行业研报的情感词词典，对研报文本的积极情感词与消极情感词进行统计，最终统计结果呈现整体平均占比如下所示：

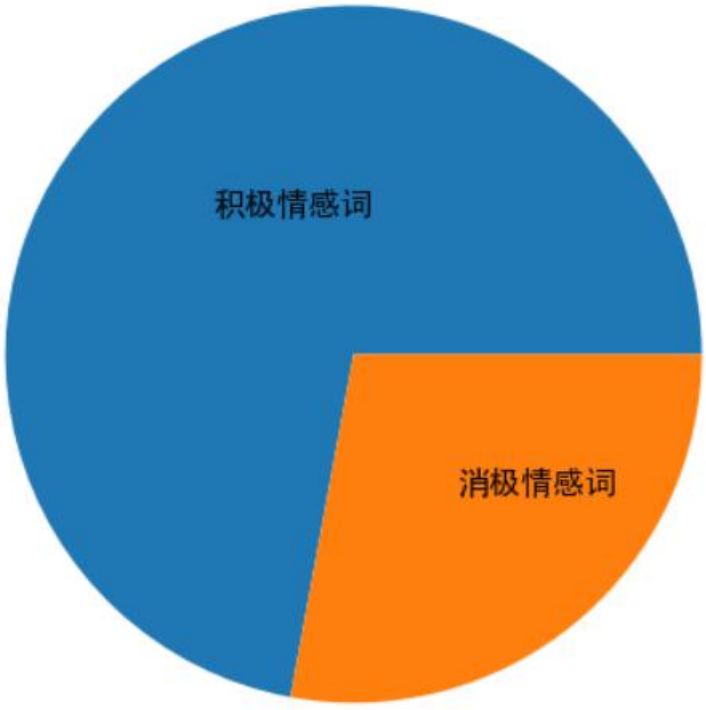

Figure 14 积极情感词与消极情感词相对占比

从图中我们可以看出，股市行业研报以积极情感词为主，消极情感词相对较少。相关论文研究表明，股市相关文章内容存在严重的春秋笔法，即报喜不报忧，通常情况下认为积极情感词占文章主导为正常现象。

附录二：因子说明

1. CCI（顺势指标）

①指标含义

CCI 指标专门测量股价、外汇或者贵金属交易是否已超出常态分布范围。属于超买超卖类指标中较特殊的一种。CCI 指标是根据统计学原理，引进价格与固定期间的股价平均区间的偏离程度的概念，强调股价平均绝对偏差在股市技术分析中的重要性，是一种比较独特的技术指标。

$$\mathrm{CCI} = \frac{1}{0.015}\frac{TP - MA}{MD}$$

②运用原理：

CCI 指标的运行区间也分为三类：+100 以上为超买区，—100 以下为超卖区，+100 到—100 之间为震荡区：

a. 当 CCI 指标曲线从上向下突破+100 线而重新进入常态区间时，表明市场价格的上涨阶段可能结束，将进入一个比较长时间的震荡整理阶段，应及时平多做空。

b. CCI 指标曲线从下向上突破-100 线而重新进入常态区间时，表明市场价格的探底阶段可能结束，有可能进入一个盘整阶段，可以逢低少量做多。

c. 当 CCI 指标曲线从上向下突破-100 线而进入另一个非常态区间（超卖区）时，表明市场价格的弱势状态已经形成，将进入一个比较长的寻底过程，可以持有空单等待更高利润。

d. CCI 指标曲线从下向上突破+100 线而进入非常态区间(超买区)时，表明市场价格已经脱离常态而进入强势状态，如果伴随较大的市场交投，应及时介入成功率将很大。

2. TAPI（指数点成交值）

① 指标含义

TAPI 指标是根据股票的每日成交值与指数间的关系，来反映股市买气的强弱程度及未来股价展望的技术指标。需要注意的是，TAPI 指标必须与其它指标结合研判，不能单独作用。

$$TAPI = AMOUNT + D$$

② 运用原理

a. 上涨过程，在股价的明显转折处，若 TAPI 值异常缩小，是向下反转讯号，应逢高卖出；连续下跌中，在股价明显转折处，若 TAPI 值异常放大，是向上反转讯号，可逢低买进。

b. 发生背离现象。即指数上涨，TAPI 值下降，此为卖出讯号，可逢高卖出；反之，为买进信号。

3. MTM（动量指标）

① 指标含义

动量指标从股票的恒速原理出发，考察股价的涨跌速度，以股价涨跌速度的变化分析股价趋势的指标。动量指数以分析股价波动的速度为目的，研究股价在波动过程中各种加速，减速，惯性作用以及股价由静到动或由动转静的现象。动量指数的理论基础是价格和供需量的关系，股价的涨幅随着时间，必须日渐缩小，变化的速度力量慢慢减缓，行情则可反转。反之，下跌亦然。

$$\text{MTM}=C - CN$$

② 运用原理

a. MTM 由上向下跌破中心线时为卖出时机，相反，MTM 由下向上突破中心线时为买进时机。

b. 若股价与 MTM 在低位同步上升，显示短期将有反弹行情；若股价与 MTM 在高位同步下降，则显示短期可能出现股价回落。

4. VMA(变异平均线)

① 指标含义

变异平均线则是用每日的开盘价、收盘价、最高价和最低价相加后除以 4 得出的数据计算平均线。即 $VMA = HF$的M日简单移动平均线

② 运用原理

a. 股价高于 VMA，视为强势；股价低于 VMA，视为弱势。

b. VMA 向上涨升，具有助涨力道；VMA 向下跌降，具有助跌力道。

5. KDJ（随机指标）

① 指标含义

随机指标 KDJ 是以最高价、最低价及收盘价为基本数据进行计算，得出的 K 值、D 值和 J 值分别在指标的坐标上形成的一个点，连接无数个这样的点位，就形成一个完整的、能反映价格波动趋势的 KDJ 指标。它主要是利用价格波动的真实波幅来反映价格走势的强弱和超买超卖现象，在价格尚未上升或下降之前发出买卖信号的一种技术工具。

要选择周期（n 日、n 周等），再计算当天的未成熟随机值（即 RSV 值），然后再计算 K 值、D 值、J 值等。

RSV 的计算公式：

$$RSV = \frac{C - L_n}{H_n - L_n} \times 100$$

其中，L_n 为之前 n 日内的最低价，H_n 为之前 n 日之内的最高价。

计算 K_i：

$$K_i = \frac{2}{3} K_{i-1} + \frac{1}{3} RSV_i$$

K_i, RSV_i 分别表示某一天当天的 K 值和 RSV 值。

计算 D_i：

$$D_i = \frac{2}{3} D_{i-1} + \frac{1}{3} K_i$$

D_i, K_i 分别表示当天的 D 值和 K 值。

计算 J 值：

$$J_i = 3K_i - 2D_i$$

② 运用原理

a. K 与 D 值永远介于 0 到 100 之间。D 大于 80 时，行情呈现超买现象。D

小于 20 时，行情呈现超卖现象。

b. 上涨趋势中，K 值大于 D 值，K 线向上突破 D 线时，为买进信号。下跌趋势中，K 值小于 D 值，K 线向下跌破 D 线时，为卖出信号。

5. OBV(能量潮)

①指标含义

能量潮是将成交量数量化，制成趋势线，配合股价趋势线，从价格的变动及成交量的增减关系，推测市场气氛。其主要理论基础是市场价格的变化必须有成交量的配合，股价的波动与成交量的扩大或萎缩有密切的关连。通常股价上升所需的成交量总是较大；下跌时，则成交量可能放大，也可能较小。价格升降而成交量不相应升降，则市场价格的变动难以为继。

$$OBV = [(close - low) - (high - close)] \div (high - low) \times volume$$

②运用原理

a. 当股价上升而 OBV 线下降，表示买盘无力，股价可能会回跌。

b. 股价下降时而 OBV 线上升，表示买盘旺盛，逢低接手强股，股价可能会止跌回升。

c. OBV 线缓慢上升，表示买气逐渐加强，为买进信号。

d. OBV 线急速上升时，表示力量将用尽为卖出信号。

附录三：关于 R^2 结果的讨论

最开始采用拟合优度（R^2）作为模型的评价指标，这样可以与其他学者的结果进行比较。训练完成后分别计算测试集中每支股票的拟合优度，取平均作为最终评价结果。由于拟合优度的取值范围为(-∞,1]，且有效值仅能取(0,1]，因此少量异常值就能使得最终结果异常。且在一次实验结果中有大量股票的取值都小于 0，极端最值为-2321.7734085731786。将几支效果极差的股票预测结果与真实结

果可视化，结果如下。

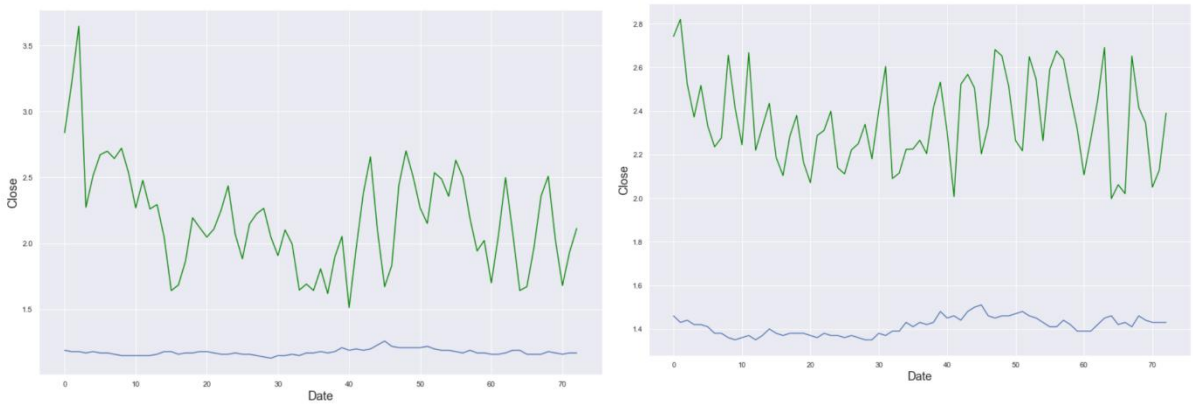

Figure 15-16 预测情况与真实情况对比图

我们猜测原因如下：原始数据中，特征由三部分组成，分别是交易因子数据、历史交易数据与文本特征。文本特征由两类研报数据处理而得，分别是个股研报与行业研报。个股研报的影响施加在这个研报研究的个股上，行业研报的影响施加在这个行业的所有个股上。文本特征仅在当日出现相关研报的情况下有值，其他时候都为 0。总的来看在文本特征部分数据非常稀疏。坚持这样处理的原因是我们把没有文本信息时的特征当作正常情况，而出现文本特征时就能额外考虑这些信息，让这些信息对预测过程产生较大影响。但对于一些股价波动不大、关注者不多的小企业的股票（我们在实验中认为平均收盘价小于 2.5 元的股票为此类股票）来说，行业研报的信息并不会对其股价产生较大影响。因此，现实情况中这类股票的股价是稳定维持在一个很小的水平上，但因为额外的行业研报信息的加入，预测值隔一段时间就突变一次。波动幅度不大的历史交易数据无法提供很多信息，更放大了行业研报的影响。

选取拟合优度小于-500 的同一行业的三支股票，将预测结果反映在一张图上。

Figure 17 预测情况与真实情况对比图

可以发现虽然是三支不同的股票，且真实值都在 1~2 之间波动，但预测值却有相同的剧烈变化的趋势。这正是每隔一段时间就加入的行业研报信息的影响。

进一步进行关于 R^2 评价指标选取的讨论。本文发现由于拟合优度是预测回归问题中唯一一个可以在不同数据集上互相比较的指标，因此我们最开始使用拟合优度来评价我们的结果。