分类号_____　　密级 _____
U D C _____　　编号 10741

硕 士 学 位 论 文

论文题目 基于文本挖掘和机器学习算法的股票投资研究

研 究 生 姓 名：陆航航

指导教师姓名、职称：史亚荣 教授

学 科、专 业 名 称：应用经济学 金融学

研 究 方 向：金融理论与政策

提 交 日 期：2022-06-05

独创性声明

本人声明所呈交的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，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

学位论文作者签名：陆航航 签字日期：2022.6.5

导师签名：史西荣 签字日期：2022.6.5

关于论文使用授权的说明

本人完全了解学校关于保留、使用学位论文的各项规定，同意（选择“同意”/“不同意”）以下事项：

1.学校有权保留本论文的复印件和磁盘，允许论文被查阅和借阅，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文；

2.学校有权将本人的学位论文提交至清华大学“中国学术期刊（光盘版）电子杂志社”用于出版和编入 CNKI《中国知识资源总库》或其他同类数据库，传播本学位论文的全部或部分内容。

学位论文作者签名：陆航航 签字日期：2022.6.5

导师签名：史西荣 签字日期：2022.6.5

Research on stock investment based on text mining and machine learning algorithms

Candidate : Lu HangHang

Supervisor: Shi YaRong

摘 要

作为资本市场的一个重要组成部分，股票市场在整个金融领域中有着十分重要的地位。一直以来，无论是在学术领域还是在资产管理行业，都有很多人希望能够找到一种能够正确地预测股票市场变动方向的方式。过去人们在进行股票价格走向的预测分析时，大部分是从股票市场自身的角度出发，利用证券价格的历史价格，交易量或者是和公司相关的收益指标来分析股票未来价格的变动，很少会从经济新闻、财经自媒体文章、股市评论等文本信息的角度来分析股票未来的价格变动，而随着互联网特别是移动互联网在百姓生活中的普及，越来越多的投资者开始接触到各种财经新闻，并且以此来辅助自己进行投资，因此研究这些财经文本信息对股市的影响，以及如何从这些文本信息中挖掘出有效的特征来对股票市场的走势进行判断有着重要的现实意义。

本文的研究是利用财经新闻文本数据，从文本挖掘的角度来分析财经新闻对股票价格的影响及利用从文本数据中抽取的特征来对股票市场的未来走势进行预测。与市场上常见的技术性指标、基本面指标等结构化数据不同，文本信息属于一种非结构化的数据，要想让模型能够理解这些信息并从中挖掘出一些有用的特征，必须先对收集到的财经新闻进行转换处理，本文首先运用了 Python 第三方库 jieba 来收集到的对每条新闻文本进行切词，将每条文本信息转换成一个个单词列表，再利用自然语言处理工具 doc2vec 将每条文本信息转换成向量形式，之后通过随机森林算法对数量化的文本特征进行特征选择，从而得到最终的进入模型的特征数据。在回测时，本文将股票预测定义为一个二分类问题，即用现在的文本特征来预测第二天股票市场的涨跌情况，预测标的为沪深 300 指数。最终比较了三种不同类型的机器学习模型支持向量机（SVM）、XGBoost 模型和神经网络在该分类任务上的表现情况，结果表明在将文本信息添加进模型之后，模型对标的的预测性能得到提高，这证明了文本信息的有效性，而在所有的机器学习模型当中 XGBoost 的表现最优。

关键词：文本挖掘 机器学习 量化投资 股票市场

Abstract

As an important part of the financial market, the stock market plays a important role in the entire financial field. For a long time, whether in the academic field or in the field of stock investment, there are many people who are full of great interest in the prediction of the direction of the stock market. In the past, when people predicted the trend of stocks, most of them started from the perspective of the stock market itself, using the historical price of securities prices,trading volume or company-related income indicators to analyze the future price changes of stocks, seldom from economic news, financial self-media articles, stock market reviews and other text information to analyze the future price changes of stocks. With the popularization of the Internet, especially the mobile Internet, the speed of information dissemination continues to increase. Investors can easily use computers or mobile phones to obtain financial information from the Internet to assist themselves in making investment decisions. The impact of the stock market and how to use these text information to analyze and predict the changes of stock prices have very important practical significance.

The research of this paper is based on the text information of financial news, from the perspective of text mining to analyze the impact and prediction of financial news on stock prices. Different from structured data such as technical indicators and fundamental indicators

that are common in the market, text information is unstructured data. First, the collected financial news needs to be processed. This article first uses the Python third-party library jieba to segment each news text collected, and convert each text information into a word list, and next use the natural language processing tool Doc2vec to represent the text information in the form of a vector, and then use the random forest algorithm to perform feature selection on the quantified text features, so as to obtain the final features of the model. This paper defines stock forecasting as a two-class model, that is, using the current text features to predict the rise and fall of the stock market the next day, and the forecast target is the CSI 300 Index. Finally, the performance of three machine learning models, support vector machine, XGBoost, and neural network on the classification task was compared. The results show that after adding text information into the model, the model's target prediction performance is improved, which proves that the text information effectiveness. Among all machine learning models, XGBoost performs the best.

Keywords：Text mining; Machine learning；Quantitative investment；Stock market

目　录

1 绪　论

1.1 研究背景与意义

1.1.1 研究背景

股票市场具有很高的流动性，相比于其它经济指标，其反应速度更加迅速，所以股票市场常被当作是指示一个国家经济运状况的晴雨表，同时股票市场在整个资本市场中有着非常重要的地位。长久以来，国内外有大量的学者和投资者专注于分析股票市场内在的运作机制，希望能够找到一种行之有效的方式来对股票市场价格的变动进行解释，进而能够预测股票价格未来变动的方向。传统的股票预测方法可以分为两种：第一种是技术分析，技术投资者认为市场上的所有信息比如国家政策、国际环境、公司经营状态甚至是投资的情绪都反映在了股票的价格当中，所以投资者不需要知道价格以外的信息，同时，技术投资者认为历史会重复发生，所以可以通过观察证券价格变动形成曲线来指导自己进行投资，当 K 线图出现某些特殊的形态时，投资者就会根据信号进行买卖交易。为了能够更加直观地分析 K 线的状态，或者是能够让电脑通过某些数量化的特征知道此时 K 线图的状态，一些学者和专业的投资者们利用股票的数据比如最近几分钟、几天、几周甚至是几个月的开盘价、收盘价成交量等数据构造出了像相对强弱指标（RSI）、移动均线（MA）、平滑异同平均线（MACD）、能量潮（OBV）等技术性指标，通过监测对应指标的变动来指导自己的投资，通常技术投资者关注的股票价格的短期波动，希望通过快速地低买高卖来获得资本增值。第二种是基本面分析，基本面投资者在投资某个证券时不会太在意投资标的过去的价格变动情况，相反他们会将更多的时间和精力用于对国家政策、行业发展趋势、公司经营情况等基本面的分析上面，在经过各方面细致的分析之后才会选择自己认为成长性最高的投资标的，基本面投资者通常更加稳健，交易频率也更低，与利用市场短期波动来赚取差价的技术性投资者不同，基本面投资者更加希望选中一家未来能够不断成长的企业进行投资，然后在公司成长后获得股票价格上涨的收益，所以相对来说其投资周期更长。不管是技术分析还是基本面分析都有一个显著的特点，

即所使用的数据都是结构化的数据，这些数据容易获得，也更好进行分析，在金融行业，针对这些结构化的数据已经有了大量的研究，也形成了一些成熟的理论。但是大量的研究表明，在这些我们能够获取到的结构化数据之外，还有大量的能够影响股票走势的信息并非以数字的形式呈现，而是以文本的形式存在。随着我国互联网特别是移动互联网技术的发展，我国网民规模有了很大程度的提升，根据互联网信息中心发布的《第 47 次中国互联网络发展状况统计报告》中的数据显示，到 2020 年 12 月底，中国的网民数量上升到了 9.89 亿，互联网在全国百姓中的普及率已经达到了 70.4%，正是由于我国互联网的普及，信息的传播速度不断提升，现在位于不同地区的投资者们可以轻易地利用电脑或者手机从互联上获得财经信息。和以往受到条件限制不同，现在的投资者能够接触到更多的新闻媒体信息，这包括专业财经媒体所做的经济走势分析，还有一些资深自媒体所做的技术面或者是基本面分析，这些分析在互联网中广为流传，越来越多的投资者开始根据这些信息来辅助自己进行投资。由此可见，文本信息在股票市场的作用也越来越大，因此从各种文本数据挖掘出能够预测股票走势的信息成为了当今金融投资领域的研究热点。

另一方面，传统的股价预测模型以时间序列分析为基础，这类模型会假设各种数据是平稳和低噪声的，数据间的关系也是线性的，但是复杂的金融市场很难满足这些假设，因此传统的模型很难真正挖掘出各个因子与股价变动之间的关系。随着以 GPU 为代表的各种硬件算力的提升和以 BP（back propagation）为代表的优化算法的改进，机器学习算法开始在无人驾驶、自然语言处理、图像识别等领域绽放光彩。而在金融领域，机器学习也在很多方面发挥了重要的作用，比如一些勇于创新的银行机构很早就将机器学习用于欺诈检测、各大资产管理公司利用机器学习研发智能投资顾问、互联网金融公司利用大数据和机器学习算法来判断新客户的信贷状况，从而给出最合适的授信额度和利率大小，大大加快了网络贷款的速度。机器学习之所以能发挥这么大的作用，主要是因为机器学习算法可以在不用对数据的分布进行任何假设的前提下，挖掘各个特征之间存在的非线性关。显然，文本挖掘技术和机器学习算法的兴起为金融量化投资的研究带来了新的突破口。

1.1.2 研究意义

相比于国外发达的量化投资交易，我国的量化投资还没有形成主流。通过将文本挖掘和机器学习方法引入量化投资，有助于形成更加稳定可信交易策略，从而吸引投资者的注意，促进我国量化交易市场的发展。

本文的理论研究意义主要有以三个方面：丰富了机器学习在经济学领域的相关研究，随着人工智能的火热，学术界开始探讨机器学习在各个领域的研究运用，希望通过这些技术来减少人力成本，提高产出效率，但总体而言，金融学界对这类工具的使用还处于探索阶段，一些传统金融机构对这些技术的接纳程度还相对较低，同时在学术界相关研究比较缺乏，本文提供了一个利用机器学习研究金融问题的典型案例，为人工智能在金融领域的研究提供了一定参考；第二，我国量化投资起步较晚，虽然有不少投资机构在对量化投资进行研究，但是相应的研究成果依旧较少，在大数据时代，能够对具有复杂关系的相关变量进行建模的机器学习方式一定会成为将来金融研究的主流之一，我国的金融发展想达到发达国家的水平，必须在在机器学习方面进行仔细研究，本文将机器学习和量化投资理念相结合有助于丰富和补充相关文献。第三，本文利用爬虫技术从网络上获取相关的文本信息，然后用文本挖掘提取特征，扩展金融学特别是行为金融学的研究渠道。

本选题的研究还具有一定的实践意义。一是各大投资公司在传统的量化投资套利空间不断缩小的情况下，急需利用更好的技术开发新的投资策略，而机器学习为资产管理公司带来了新的思路，如果成功将机器学习运用到各大投资的量化投资策略中，将有助于形成更加有效，更加稳定的交易策略。二是促进我国量化投资市场的发展，据统计，量化投资在一些发达国家的金融投资领域的运用占比达到了一半以上，但是量化投资在我国的占比不到十分之一，通过将分析文本挖掘和机器学习在量化投资中的运用，将增加量化投资在市场交易中的地位。三是使用文本分挖掘机器学习算法能够让金融机构在不需要增加人力的情况下对市场上存在的信息进行更有效的挖掘和分析，这将使得信息在股市中流动的效率提升，从而提高我国资本市场的效率。

1.2 文献综述

1.2.1 文本信息和资本市场的关系研究

财经新闻媒体包含的信息十分丰富，包括了国际环境、行业政策、上市公司财务等各种信息，在这些信息流入市场之后，接触到这些信息的投资者会受到媒体当中某些观点或者是消息的影响，然后在这些文字信息的影响下，投资者的投资态度可能发生改变，进而会影响整个证券市场的波动。大量研究证实，各种新闻媒体的消息会对证券价格的波动产生影响，游家兴等（2012）利用文本挖掘技术从关注水平、新闻基调、曝光水平等几个角度创造了能够反映媒体情绪的指标体系，在此基础上研究了股票价格变动和媒体情绪的联系，研究结果表明，新闻媒体的情绪不同，对股票资产价格变动的影响存在一定程度的不对称性，积极的情绪会驱使股票价格上涨，从而形成价格泡沫，而且当标的公司的信息公开程度越低时，新闻媒体情绪对股票价格的影响效果越明显；牛枫等（2017）从媒体新闻的监督和媒体新闻的关注两个维度分析了新闻媒体的发布对上市公司 IPO 定价的影响，实证结果表明当一家公司在进行公开上市时，如果其获得的媒体关注度较高，则会获得更高的 IPO 定价，即两者之间存在着正相关关系，而媒体监督则是和 IPO 之间存在着负相关关系，同时作者也证明了新闻媒体报道对 IPO 价格的影响主要是源于对报价机构报价最终报价水平的影响；龙文等（2019）利用主题模型对市场上宏观财经新闻的话题分布情况进行分析，证实了我国 A 股市场上存在“媒体效应”，并在此基础上研究了新闻主题与投资收益率的关系，结果表明对财经新闻的主题分布情况的分析可以为股票投资带来明显的超额收益；吕华揆等（2021）利用 word2vec+K-Means 方法对收集到的新闻文本进行了聚类分析，之后运用 VAR 模型分析了各种类型的新闻是如何造成股市的变动，实证结果表明，不同种类的媒体情绪效应和信息效应可以对股票市场的成交量、震动幅度和收益率产生影响，但是各种新闻类型对股票市场影响的侧重点各不相同；姜富伟等（2021）利用 word2vec 算法构建了中文金融情感词典，利用该词典和大量的财经新闻文本计算出了媒体情绪指标，该指标能够对未来的股市进行预测，而且表现良好。

在国外，Mittermayer（2004）展示了一种根据新闻分类进行投资的交易系统（NwesCATS），该系统分为三部分，第一部分是利用文本挖掘技术来从最近发布的新闻中提取信息，第二部分利用机器学习模型将处理后的信息进行分类，第三部分则利用分类之后的信息进行投资决策，实证结果表明该分类模型可以能为预测股票市场的走势提供额外信息；Blaufus（2019）分析了德国 176 家公司从 2003 年到 2016 年的税收新闻与公司股价变动的关系，发现当出现合理避税的相关新闻时，公司股价呈现下降趋势，而新闻中出现非法逃税相关内容时，公司股票在未来一段时间的表现较差；Tobias（2020）通过分析“中美贸易战”期间美国总统特朗普在推特上发布的 3200 条消息，证实了政治新闻和股票走势之间的关系，结果表明贸易战的消息会让标普 500 指数下跌，同时使得美国恐慌指数 VIX 上涨，同时格兰杰因果检验表明这种影响是单向的，即新闻影响了股市和 VIX 指数的变动；Salisu（2021）根据感染人数和死亡例数选出了受新冠肺炎影响程度最大的 20 个国家，然后分析了疫情期间股票走势和医疗健康新闻的相关性，结果表明在预测模型中添加和健康新闻的因素可以大大提高预测的精度，证明了在疫情背景下健康新闻的发布会对股市的波动产生影响。

1.2.2 机器学习在预测股市方面的研究

在利用机器学习构建量化投资模型方面，我国的学者也做了许多相关研究：王燕、郭元凯（2019）通过网格参数寻优算法对 XGBoost 机器学习模型进行了参数优化，在此基础上构建出了 GS-XGBoost 金融投资预测模型，然后经过实验对比后，发现与 XGBoost 元模型、梯度促进决策树模型以及支持向量机模型相比，GS-XGBoost 模型在均方误差（MSE）、平均绝对值误差(MAE)与均方根误差(RMSE)等多种评价指标上都能获得更好的预测表现；李斌，邵新月，李玥阳（2019）将套索回归算法、岭回归算法、长短期记忆模型、支持向量机（SVM）、集成神经网络等十二种常用的机器学习用于股票价格变动的预测，并在预测的基础上构造出了对应的投资组合，回测结果表明相比于传统的线性模型，能够寻找非线性关系的机器学习算法能够更好地对对各种异象因子进行识别，挖掘出更大的超额收益；赵红蕊，薛雷（2020）首先将长短期记忆模型和卷积神经网络相进行结合，之后在这两个模型的基础上又引入注意力机制，形成了一个深度复合模型，该模

型是一种能够进行端到端训练的深度神经网络架构，在对样本进行训练时，数据首先通过长短期记忆模型模型来提取金融序列中的时序特征，然后再通过卷积神经网络来对数据中的深层特征进行挖掘，最后通过在网络结构中添加注意力机制，以此进一步提高模型的性能，这种结构有效地提升了模型挖掘有效特征的能力，从而能够更加准确地预测股票价格的变动；张倩玉，严冬梅，韩佳彤（2020）针对股票价格预测存在的非平稳和非线性问题，将深度神经网络与分解算法相结合，该模型通过能够自适应噪声的完整集成经验模态分解（CEEMDAN）算法获得证券价格时间维度上的信息，在这之后又通过注意力机制挖掘输入特征参数的权重，然后使用门控循环单元模型来对投资标的的价格变动进行预测，回测的结果表明，该混合模型在苹果、贵州茅台等国内外四家公司的股票价格和上证指数上的表现要比 RNN、LSTM 等模型的预测误差要小。

在国外，Yoshihara(2014)将循环神经网络（RNN）用于股市预测，RNN 是一种专门用于处理时间序列的机器学习模型，该模型在预测未来的趋势时能够对过去的相关序列进行考虑，其在自然语言处理方面已经取得了很大的成就，作者将该模型引入金融时间序列的分析，并在日经指数上进行了测试，证明了该方式的实用性，但是 RNN 由于优化问题，不能记住过长的时间序列，所以该方法还有待改进；Usmani(2016)等人将油价变动，金价银价变动，外汇变动等作为特征，然后用多层感知机，径向基函数，支持向量机等机器学习方法对卡拉奇证交所的股指收盘价进行预测，证明了 KSE-100 指数能够被机器学算法预测，而在所使用的机器学习算法中，多层感知机的表现要优于其它算法；Nelson(2017)将长短期记忆模型（LSTM）用于处理金融时间序列，该模型通过一个特殊单元“gates”，来缓解反向传播中梯度随着序列长度增加而消失的缺陷，该方法能够在一定程度上缓解 RNN 只能对较短时间的数据情况进行记忆的缺点，通过结合更长范围的时间序列来判断证券价格未来的涨跌情况，作者在巴西股市中对该方法进行了测试，结果表明 LSTM 在预测未来一段时间内涨跌的准确率达到了 55.9%。值得注意的是，虽然 LSTM 神经网络在一定程度上缓解了循环神经网络记忆较短的缺点，但是并未将其完全克服；Nikou（2019）等人将深度学习，随机森林，神经网络，支持向量回归机等机器学习算法应用于英国股市的股指预测，发现深度学习的预测效果最好，支持向量机效果次之；Nobre(2019)将 XGBoost 算法用于股指期货

预测，该方法是一种性能十分优越的分类算法，在很多国际赛事中都能看到它的身影，作者将其用于韩国指数期货的预测，并将其和 LSTM 和传统的自回归时间序列处理方法进行比较，实证结果表明，XGBoost 算法在判断涨跌方面效果更优。

1.2.3 文献评述

可以看出，大量的学者对文本信息和资本市场的关系做了大量的研究，证实了文本信息中包含着大量能够预测未来股票收益的信息，但是以往的研究主要集中于从文本信息中提取比如情绪，关注度，监管程度，话题分布等特定因子来研究未来资本市场的变动，但是却对直接利用完整的文本信息进行投资决策却鲜有研究，本文尝试将新闻文本中所隐含的信息进行特征抽取，来帮助进行股票波动的预测。另外，大量的学者证实了机器学习算法在股票预测模型中的高效性，因此，本文将不再使用传统的线性回归来进行建模而是选择支持向量机、XGBoost、神经网络三种机器学习模型来进行股票预测。

1.3 研究框架与思路

本文是对财经新闻和股票市场变动之间的关系进行分析，首先要做的是对文本信息的处理，在将文本信息利用 doc2vec 模型进行量化以后，再使用随机森林算法对相关特征进行选择。最后利用三种不同的机器学习算法，根据提取到的文本特征来进行股票预测。

本文的研究结构设置如下，第一章为绪论，这一章主要阐述本文的研究背景和研究意义，还有对国内外的相关研究进行文献综述。第二章是对本文所应用的相关理论与算法进行介绍，主要介绍了支持向量机（SVM）、XGBoost 模型、神经网络等三种不同类型的机器学习算法，还有进行文本特征提取所用到的模型。第三章，数据处理，介绍了本文所使用的文本数据和股指数据的来源，还有具体的处理步骤，为接下来模型的训练做准备。第四章，利用机器学习算法和文本特征对股指涨跌进行预测，展示了实证分析的结果。第五章为结论与启示，对全文进行总结进而得出启示与建议。

1.4 本文的创新之处

（1）目前大部分关于股票市场预测的所做的分析，都是利用股票市场上所累积的历史信息，按照时间序列的方式进行研究，而本文则是现对大量的财经新闻本文进行文本挖掘，然后用这些抽取出来的特征来对证券价格的未来走势进行判断，这为金融投资研究提供了一种新的思路。

（2）针对词向量过高的问题，引入随机森林算法对高维特征进行筛选，选出表现最好的因子，这一方面解决了数据量过多而造成训练难度加大问题，另一方面减少了用于预测的特征，从而缓解了机器学习算法中经常出现的过拟合问题，提高模型的预测性能。

（3）本文引入了支持向量机、XGBoost、神经网络等机器学习算法来构造模型预测股市波动，以传统的回归方法不同，这些机器学习算法能够挖掘出各种因子间的非线性关系，提高对数据的拟合能力。

2 相关理论与算法

2.1 行为金融理论

2.1.1 噪声交易者

噪声交易者这一概念最早是由 Black 于 1986 年提出，它指的是市场上大量存在着的无法获取内部消息，而将大量的噪声当成有效信息来进行交易的投资者。噪声的来源可以分为两种情况：首先从客观上来说噪声指的是市场上与证券内部价值内有关系的信息，比如公司内部编制的具有欺骗性的财务报告，或者是一些机构通过违法手段造成的市场异常波动；从主观上来说是市场上的投资者由于自身知识的局限而对获取到的信息产生了错误的解读，或者是自身能力的不足而导致了认知上的偏差，从而做出的非理性投资决策。

产生噪声交易者的原因主要是来自以下两个方面：

（1）信息不对称

投资者的最终目标是使得自身的利益最大化，而产生不同投资决策的主要原因是各个投资者能够能够从市场上获取到的信息不同，而得到的信息的不同在很大程度上是由取得信息的约束条件所决定的。

根据每个人获取信息的约束条件的不同，投资者可以被分成“有信息投资者”与“无信息投资者”。一些投资者为了博取更大的收益在一开始就投入了大量的资金甚至是背负了高额债务，这些投资者为了能够成功获利或者是避免大额损失，在投资时会主动在市场上寻找大量和证券价值有关的信息，甚至还会想办法介入公司的治理过程当中，种种原因使得这一部分投资者成为了“有信息投资者”。而“有信息投资者”收集到的信息会通过各种渠道在市场上进行传播，即使有些投资者不去主动收集信息，最后还是会接收到一些消息，这些自身没有意愿进行信息收集的投资者被称为“无消息投资者”。

由于“无消息投资者”最终在市场上收集到的信息量相对较少，而且这些信息的来源没有“有消息投资者”那么正规，这些投资者在进行投资决策时更加偏向于使用市场上的“小道消息”，甚至有些投资者会直接模仿一些“投资专家”所做的交易，最终形成了噪声交易。

（2）认知偏差

认知偏差指的是市场上的投资者由于自身所处环境的限制而导致对于特定信息的处理能力存在一定的局限性，最终使得每个人所作出的决策存在一定的偏差。从心理学的角度来进行研究，所有的投资决策都是投资者对收集到的相关信息进行加工和分析的过程，在这个过程当中很有可能产生三种认知偏差：过度自信偏差、保守主义偏差和从众性偏差。这些投资者在处理信息的过程当中没能正确识别出股票市场上的错误信息，最终让自己做出了一些错误的决策，产生了噪声交易。

2.1.2 有效市场假说及挑战

有效市场假说由金融学家 Fama 在 1965 年正式提出，它指的是在整个金融市场当中，所有的投资者都是理性的，所有投资者的目标都是追求自身利润的最大化，每位投资者都会在这一市场当中进行积极地竞争，都想通过正确地预测股票的价格走势来获取大量的收益。在这一市场当中证券的价格完全反映了市场上所有可以获得的信息，从而使得任何投资分析都无效化。

有效市场假说认为金融市场可以被划分为三种不同的状态，分别是弱有效市场、半强式有效市场与强有效市场，这三种状态中，股票价格所能反应的信息量依次增加。在弱式有效市场假说当中，我们认为此时的证券价格已经完全消化了所有与证券价格有关的历史信息，这包括股票的成交量、融资金额、股票过去的成交价等信息，由于这些信息都是公开透明的，投资者会针对这些信息进行证券的买卖，从而使股价得到调整，此时任何图表或者是技术分析都将失效，只有基本面分析或者是内幕交易才能获取超额利润。半强式有效市场假说则认为现在的证券价格不仅包括了过去的历史信息还有市场上出现的关于公司运营前景的相关信息，比如公司的管理状况或者是对外公开披露的财务信息，在半强式市场中，不仅技术分析失效，基本面分析也无法发挥作用，只有内幕交易者才有可能获得超额收益。而强式有效市场假说认为当前的证券价格完全反映了关于该公司的所有信息，这些信息不仅包括已经在市场上公开披露信息还包括还未经过市场披露的内部效息，在这样的市场当中，即使是内幕投资者也无法获得任何的额外收益，任何投资者最终所能获得的边际市场价值都是零。

随着时代的不断发展，股票市场也在不断地进行迭代，在后续的研究中，有效市场假说的前提条件和实际情况并不相符，这主要表现在以下几个方面：

（1）投资者并非总是理性的

在有效市场假说当中有一个非常重要的前提假设，即这个市场上的金融投资者都具有足够的理性，并且能够在短时间内对市场上的所有信息做出正确合理的反应。但是在现实生活中，每个人由于所处的环境不同，接触到的事物不同，所有每个人或多火烧会形成自己的习惯还有偏好，这导致每个人在面对大量形式各异的数据时很难保持理性。而且“理性经济人”的这一假设一开始就是建立在信息对称之一假设之上，但是每个人的认知水平或者是获取信息的能力很可能存在差异，使得场上的投资者存在信息不对称的问题，这些差异和问题的存在最终很有可能让投资者们做出完全不同的选择。

（2）市场交易并不是随机的

有效市场假说假定市场上的投资者都是相互独立的，市场上发生的交易行为也都是随机的。但是，从后来的研究中发现，金融市场上发生的交易并不都是随机的，任何人之间经常存在着非常复杂的非线性关系，常常会出现因为其中一个人的行为引发“蝴蝶效应”，最终造成了大量的非理性交易行为，这些非理性的交易行为往往会形成一股力量导致证券价格往一个方向移动，有可能在短时间内无法得到修正，甚至会在一个很长的时间周期内处于一种偏离的状态。股票市场上的“羊群效应”就是一个非常明显的例子，大家在受到市场上某种情绪或者是是某种事件的发生后就会产生集体性行为。

（3）市场套利的有限性

有效市场假说认为，只要有一个十分健全、有效的市场，该市场上的所有投资者都能进行无风险和无成本的套利，那么任何时候证券的价格发生偏离都会有投资者通过这种套利来使得证券的价格得到修正，使得资产的价格回归到正常水平。这一观点实际是建立在两个假设之上：一是一旦市场上出现新的套利机会，就会有投资者进行套利交易从而获取利润；而是套利行为的发生不存在任何限制，所以资产价格会很快得到调整。

但是在现实生活中存在着大量的因素阻碍着投资者的套利行为，比如投资者自身存在认知局限性，无法获得套利信息，又或者市场上的规则限制了投资者

的投资行为，从而会影响资产价格回归。

2.1.3 有限关注度理论

传统金融分析理论经常认为金融市场是有效率的，股票的价格是对市场上所有信息的充分反应。但是，20 世纪 80 年代市场上出现了大量传统金融理论无法解释的异常现象。其中很大一部分原因就是现实生活中的人很难满足“理性经济人”的前提假设，在真实的市场环境当中，金融市场种的投资者的认知水平、处理信息的能力还有注意力都是有限的，正是因为这些限制的存在，使得有效市场理论在后期的发展中成为了其它金融市场理论的参照系。后来的学者为了对市场上存在的各种异象进行解释，开始将一些认知心理学尤其是行为科学的相关研究用于对投资者行为进行分析，这种研究方式逐渐形成了现在的行为金融流派，这为传统金融理论的创新和发展做出了巨大的贡献。在行为金融学中，投资者的有限关注十个十分引人瞩目的理论，有限关注度理论认为市场上出现的大多数异象比如羊群效应，小公司效应，周末效应等都是因为人的注意力是有限的而并非是无限的。正是由于人的注意力是有限的，在相关新闻报道发布后，会针对新闻内容进行股票操作，而忽视了其它信息，从而导致了股票价格的波动。

2.1.4 媒体效应

随着互联网的普及，现在投资者们获取财经新闻的成本相比以往有了很大幅度的下降，他们可以通过电脑或者是手机及时地获取到最新的财经新闻，而根据传播学当中的媒体议程设置理论，虽然新闻媒体的报道不能直接改变观众对于某件事情的看法，却可以通过信息的传播和相关议题的设置来引导人们对事件的关注程度和思考事件的顺序，继而会影响他们的投资决策。根据行为金融学的相关研究，我们发现，投资者对与市场上流通的信息的处理能力是有限的，在接受这些信息时很容易会出现认知偏差的现象，进而导致投资者们在进行决策时的信念和偏好会形成系统性偏差，比如“议程设置偏差”。

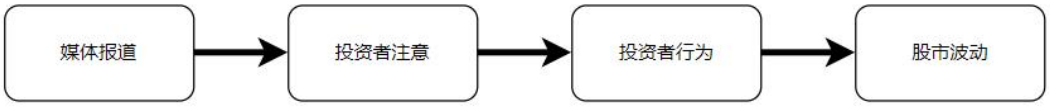

图2.1 媒体效应的传导路径

而在备受关注的股票投资领域，由于财经新闻的强大的影响力，大部分的投资者都原因将这些专业的新闻媒体当作辅助自己投资的信息来源，利用媒体发布的相关信息来构建自己的投资决策。可以看出，财经新闻在描述事件的同时，也为观众创造了事件。媒体对事件的筛选和表达，影响了投资者对所看到信息的重视程度，也就意味着，投资者最终所作出的决策会依赖于各种媒体构建出的认知环境，和其描述问提所选择的“框架”。媒体报道的框架主要包含了主体对于相关事件的选择、强调、解释和反馈，在制定框架时，新闻媒体为了追求自身利益的最大化，通常会为了迎合观众的口味，有针对性地改变整体的报道框架，过度关注当时的热点事件，期望能够获得较好地轰动效应，这也必然导致在进行新闻报道时发生有选择性的偏见，而且通过新闻媒体发布的消息很容易在市场上被广泛传播，这些偏见在市场当中不断累积，继而影响了整个市场的波动。

2.2 机器学习算法

2.2.1 支持向量机

支持向量机是由 Cortes 和 Vapnik 两位学者于 1995 年首先提出，该算法背后有着很强的理论支撑，在提出以后在学术界和工业界的受到了广泛运用。支持向量机是通过在训练过程中寻找一个能够使得样本间隔最大化的超平面来使得模型的结构风险最小化，在深度神经网络出现之前，该方法一直是监督学习的代表性算法，即使是现在，支持向量机也依旧在文字识别、图像识别、目标监测等领域有着广泛地运用。支持向量机既能用于回归任务也能用于分类任务，由于本文主要是利用文本特征来对股市未来的涨跌进行预测，所以是用支持向量机来处理分类任务。

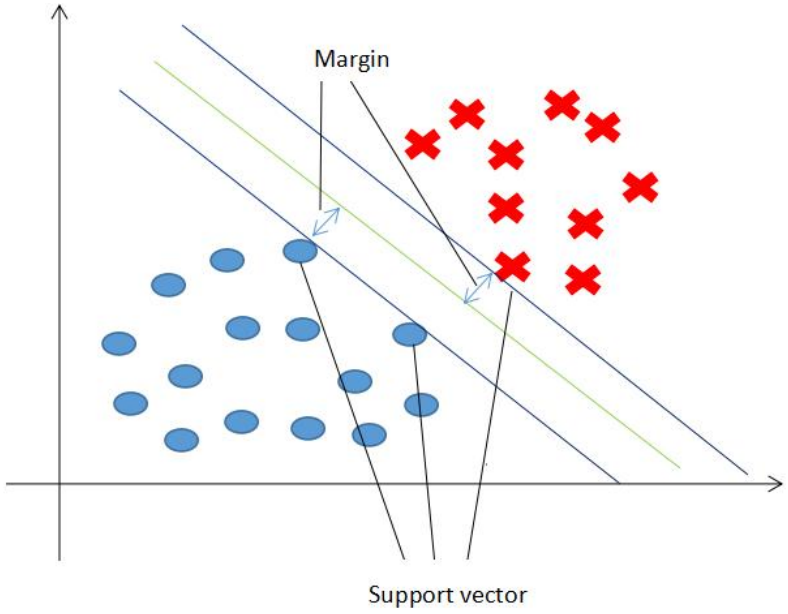

图2.1 支持向量机

如图 2.1 所示，SVM 在样本点线性可分时，会在样本点之间寻找一个超平面将所有样本点分开。SVM 的特点是对超平面的好坏有着自己的评价标准，只有距离分隔平面最近点的距离最大的那个分隔平面才是最好的超平面，因此 SVM 又被称为最大间隔分类器。

对于给定的样本集合：$D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_i, y_i)\}, x \in R^n, y \in \{-1, 1\}$，该 N 维空间中的线性判别函数表达式为 $g(x) = w \cdot x + b$，此时用于分类的超平面可以用 $w \cdot x + b = 0$ 来表示。在给定样本的情况下，我们可以通过对参数 w 和 b 进行相同比例的调节，从而让两类所有样本都可以满足$|g(x)| \geq 1$，此时，模型所得出的分类间隔为$\frac{2}{||w||}$，因此，寻找求最大间隔的目标就可以转化为让$||w||$的值最小。此时，所有满足$|g(x) = 1|$的样本点，都离模型构造出的超平面的距离最小，所有满足这一条件的样本点一起决定了最优分类面的所在位置，这些特殊的样本点又被称作支持向量。综上所述，寻找最优分类面的问题可以转化为如下优化问题：

$$\min \frac{1}{2}||W||^2 \quad s.t. \quad y_i[(W \cdot X_i) + b] \geq 1 \quad i = 1,2,3 \cdots n$$

该优化问题可以转化为：

$$\min Q(\alpha) = \frac{1}{2}\sum_{i,j=1}^{n} \alpha_i \alpha_i y_i y_i (X_i \cdot X_i) - \sum_{i=1}^{n} \alpha_i$$

$$s.t.\ \ \alpha_i \geq 0 \ \ i = 1,2,3\cdots n \ \ \sum_{i=1}^{n} y_i\alpha_i = 0$$

其中α_i是拉格朗日因子，最终可以求出判别函数为：

$$f(X) = sgn\{\sum_{i=1}^{n} \alpha_i y_i(x_i \cdot x) + b\}$$

由于对大部分数样本来说$\alpha_i = 0$，唯有与支持向量对应的$\alpha_i \neq 0$，即最优分类面的位置是由支持向量决定的。其中参数 b 为分类阈值，最后可以使用任何一个支持向量来求值。

当两类样本纠缠在一起，强行选择一个最大间隔让所有样本能够正确分类，可能导致最终选择出来的超平面在训练集上表现良好，但是泛化性较差，即出现过拟合问题，为此，我们可以在对原来的限制条件进行放松，不强求所有的样本都能被正确分类，而是给予一定的缓冲空间，从而让最终模型的泛化性能得到提升。在公式上可以表示为，在原来的方程中添加一个松弛变量$\varepsilon_i \geq 0$，使得表达式为：

$$y_i[(w \cdot x_i) + b] - 1 + \varepsilon_i \geq 0, i = 1,2,\cdots,n$$

此时我们的目标变为$(w, \varepsilon) = \frac{1}{2}||w||^2 + C(\sum_{i=1}^{n} \varepsilon_i)$最小，这表明我们是通过同时考虑间隔大小和样本分类的正确数来选择最佳的分类平面。其中 C 表示一个大于零的常数，通过改变 C 的值来对错误分类样本的惩罚力度进行调节，C 越大表明越不允许超平面将样本分类错误，C 越小，表示模型能够允许样本分类错误的样本树越多，在机器学习算法中，C 属于一种超参数，最终需要通过交叉验证来确定 C 值的大小。

在现实生活中，样本的分布经常是复杂多变的，很难满足线性可分的的条件，为挖掘样本之间的分线性关系，Vapnik 将核空间理论引入支持向量机。首先，面对线性不可分的问题，支持向量机是通过为样本添加更多特征，将样本从低维空间往高维空间进行映射，经过合适的映射，样本在高维空间可以呈现出线性可分的状态，此时就能找到能够将间隔最大的分类超平面。通常为样本添加特征可能引起计算量扩大的问题，很难进行训练，但是 Vapnik 证明我们无需将样本进行计算，只需要对相应的核函数进行低维计算，即可得到样本特征在高维空间的

关系，大大降低了计算量。在支持向量机中，常用的核函数有四种：线性核、多项式核、高斯核与 Sigmoid 核函数。

（1）线性核函数：

$$K(x,x^{'}) = x \cdot x^{'}$$

（2）多项式核函数：

$$K(x,x^{'}) = ((x \cdot x^{'}) + c)^{d}$$

（3）高斯径向基核函数：

$$K(x,x^{'}) = \exp(-\frac{||x - x^{'}||^2}{2\sigma^2})$$

（4）Sigmoid 核函数：

$$K(x,x^{'}) = \tanh(\beta_0(x \cdot x^{'} + \beta_1))$$

不同的核函数代表不同的映射关系，最终取得的超平面也不相同，除了以上提到的几种核函数，通常还需要针对不同种类的问题来构造与之对应的核函数。同样，要确定一个核函数的表现是否比其他的核函数效果更好，还需要通过交叉验证来进行判断。

2.2.2XGBoost

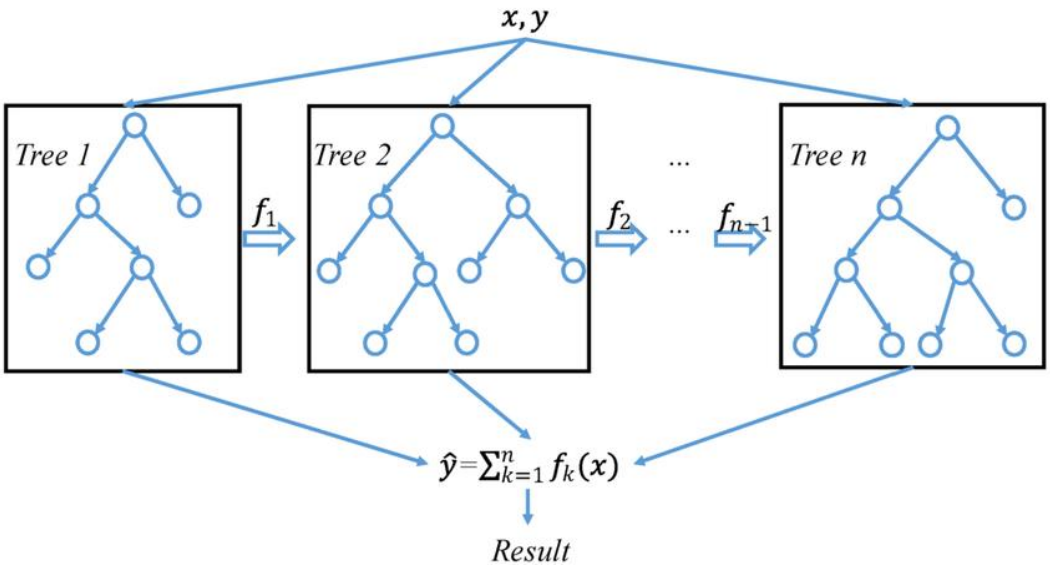

图2.2 XGBoost模型

XGboost 本质上是一种提升树模型，提升树是通过将多个弱决策树模型进行

相加来不断对分类结果进行优化的一种学习方式。Frediman 在 2001 年提出了梯度提升训练方式，该方法大大提高了模型的训练速度，梯度提升首先对目标函数在在当前模型的负梯度进行计算，然后把该值看作是模型残差值的近似。XGBoost 则是在梯度提升的基础上对模型的损失函数进行了优化，不止利用了函数的一阶信息，还利用了目标函数的二阶信息，从而使得模型的训练更加高效。

假设一个样本集 D 中有 n 条样本，这些样本的特征数量为 m，即 $D=\{(x_i,y_i)\}(|D|=n,x_i\in R^m,y_i\in R)$，此时，由 K 个基函数相加所得到的基函数可以表示为：

$$\hat{y_i}=\emptyset(x_i)=\sum_{k=1}^{K}f_k(x_i),f_k\in F$$

其中，$F=\{f(x)=w_{q(x)}\}(q:R^m\to T,w\in R^T)$表示的是有所有基函数所组成的函数空间，q 表示样本空间与叶子节点间的映射，也就是决策树的结构参数，w_i表示决策树的权重大小，T 表示决策树的叶子节点数量，每个基函数f_k都有属于自己的树结构参数 q 和表示每个叶子节点权重的 w。对于每个给定的样本点，模型会按照顺序使用 K 棵决策树规则将样本点分进各自的叶子节点当中，最后把在各棵决策树中得到的值相加即可求出对应的预测值。

（1）XGBoost 算法的目标函数如下：

$$Obj=\sum_{i=1}^{n}l(\hat{y_i},y_i)+\sum_{k=1}^{K}\Omega(f_k)$$

$$\Omega(f)=\gamma T+\frac{1}{2}\lambda||w||^2$$

XGBoost 模型的目标函数可以由经验损失和结构损失构成，一般常用的经验损失函数有两种，分别是平方损失函数与逻辑回归损失函数。其中平方损失函数的形式为：$l(y_i,\hat{y_i})=(y_i-\hat{y_i})^2$，平方损失函数用于回归预测任务的训练；逻辑回归损失函数为：$l(y_i,\hat{y}_i)=y_i\ln(1+e^{-\hat{y_i}})+(1-y_i)\ln(1+e^{-\hat{y_i}})$，该损失函数用于分类任务的训练。Ω 表示模型的正则项，正则项用于防止模型过于复杂，避免出现过拟合现象，γ，λ 为模型的超参数，它们分别用于调整模型的叶子节点的数量和叶子节点的权重大小。

（2）XGBoost 学习第 t 棵树

假设进行第 t 次迭代训练的决策树为f_t，则有：

$$\hat{y}_i^{(t)} = \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i)$$

其中，$\hat{y}_i^{(t)}$表示第在经过 t 次迭代后样本 i 所得到的预测分，$\hat{y}_i^{(t-1)}$为之前 t-1 棵决策树的预测分数，$f_t(x_i)$为模型中第 t 棵决策树的函数表达式。

由此可以得出目标函数的表达式为：

$$\mathrm{Obj}^{(t)} = \sum_{i=1}^{n} l(y_i + \hat{y}_i^{(t)}) + \sum_{i=1}^{t} \Omega(f_i)$$

$$= \sum_{i=1}^{n} l[y_i, \hat{y}_i^{(t-1)} + f_t(x_i)] + \Omega(f_t) + \mathrm{constant}$$

（3）泰勒二阶展开

对上式进行二阶泰勒展开可以得出目标函数的近似值：

$$\mathrm{Obj}^{(t)} \approx \sum_{i=1}^{n} [l(y_i, \hat{y}_i^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^{\,2}(x_i)] + \Omega(f_t) + \mathrm{constant}$$

去掉上式当中的各个常数项项，可以得出需要优化的目标函数：

$$\mathrm{Obj}^{(t)} \approx \sum_{i=1}^{n} [g_i f_t(x_i) + \frac{1}{2} h_i f_t^{\,2}(x_i)] + \Omega(f_t)$$

其中，g_i，h_i分别表示的是损失函数 L 对$\hat{y}^{(t-1)}$的一阶偏导和二阶偏导。

（4）定义决策树及其复杂度

XGBoost 模型中应用的树模型的构成主要包括叶子节点的权重向量 w 和实例叶子节点之间的映射关系 q，使用数学公式可表述为：

$$f_t(x) = w_{q(x)}, w \in R^T, q: R^d \to \{1,2,\cdots,T\}$$

决策树的复杂度Ω由其叶子节点的数量和叶子节点权重向量的 L2 范数构成：

$$\Omega（f_t） = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^{\,2}$$

（5）叶子节点归组与树结构打分

把所有进入到第 j 个节点的样本，都划划分到一个叶子节点样本集里面，即$I_i =$

$\{i|q(x_i)=j\}$，可得目标函数为：

$$\mathrm{Obj}^{(t)} \approx \sum_{i=1}^{n}[g_i f_t(x_i)+\frac{1}{2}h_i {f_t}^2(x_i)]+\Omega(f_t)$$

$$=\sum_{i=1}^{n}[g_i f_t(x_i)+\frac{1}{2}{f_t}^2(x_i)]+\gamma T+\frac{1}{2}\lambda\sum_{j=1}^{T}w_j^2$$

$$=\sum_{j=1}^{T}[(\sum_{i\in I_i}g_i)w_j+\frac{1}{2}(\sum_{i\in I_i}h_i+\lambda)w_j^2]+\gamma T$$

为了能够对上面的表达式进行简化，我们可以定义G_j和H_j来表示叶子节点 j 所包含样本的一阶偏导、二阶偏导的累加之和，它们都是常数：

$$G_j=\sum_{i\in I_j}g_i$$

$$H_j=\sum_{i\in I_j}h_i$$

即得到最终需要优化的目标函数：

$$\mathrm{Obj}^{(t)}=\sum_{j=1}^{T}[(\sum_{i\in I_I}g_i)w_j+\frac{1}{2}(\sum_{i\in I_i}h_i+\lambda){w_j}^2]+\gamma T$$

$$=\sum_{j=1}^{T}[G_j w_j+\frac{1}{2}(H_j+\lambda){w_j}^2]+\gamma T$$

对于上面的式子，由于每个叶子节点的子目标式都是完全独立的，即当各叶子节点的子目标式都取得最值时，模型最终目标函数才取得最值。假设目前树的结构已经固定，可解出各个叶子节点的权重集此时取得的最优目标值：

$$w_j^*=-\frac{G_j}{H_j+\lambda}$$

$$\mathrm{Obj}^{(t)}=-\frac{1}{2}\sum_{j=1}^{T}\frac{G_j^2}{H_j+\lambda}+\gamma T$$

2.2.3 神经网络

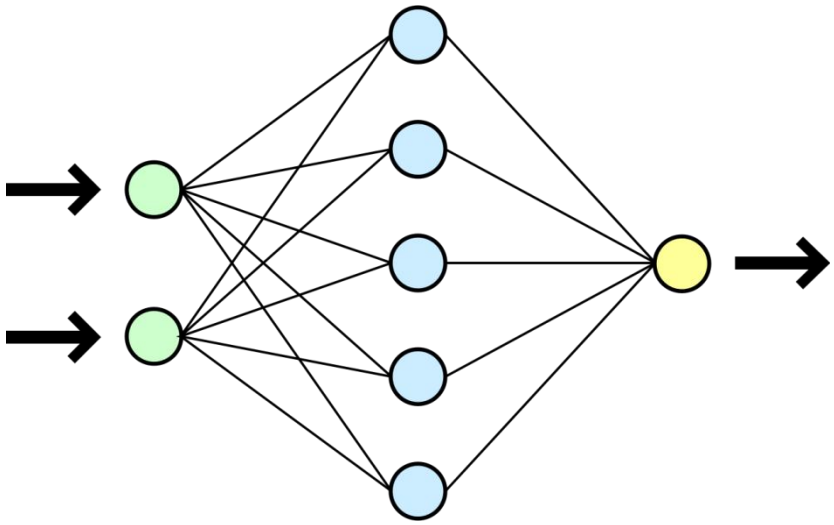

图 2.3 包含一个隐藏层的神经网络

人工神经网络是现在最为流行的一种机器学习算法，由于其强大的非线性拟合能力，还有各种优化算法的出现，使得神经网络在图像识别、语音识别甚至是无人驾驶能人工只能领域中得到广泛运用。神经网络对于在拟合非线性函数时有着非常好的效果，神经网络一般由最前端的输入层，中间的隐藏层和最终的输出层构成，在训练时，数据首先进入输入层，经过线性运算达到隐藏层，经过隐藏层的非线性转换之后会到达输出层，输出层表示模型给出的结果，将这个预测结果和真实结果进行对比分析，可得出此时的预测误差，然后通过反向传播改变参数值的大小，最终使得模型的产出和真实值之间能够有较小的误差。

对于一个由 M 个输入变量构成的向量 $x \in R^{1\times M}$，在进行如到隐藏层时得到：

$$z = xw^{T} + b_0, w \in R^{M\times M}, b_0 \in R^{1\times N}$$

其中矩阵 W 表示输入的数据道道隐藏层时所进行的线性变换。b 为每个神经元上的常数偏差。在得到前向传播的结果 z 值以后，会进入激活函数$\sigma(z)$中进行非线性转换，激活函数是神经元拟合非线性函数的核心，如果没有隐藏层所进行的非线性变换，神经网络只能拟合线性关系，而无法完成非线性关系的拟合，所以激活函数非常重要，如图 2.4 所示，常用的激活函数有 sigmoid，tanh，relu，leaky relu 等。

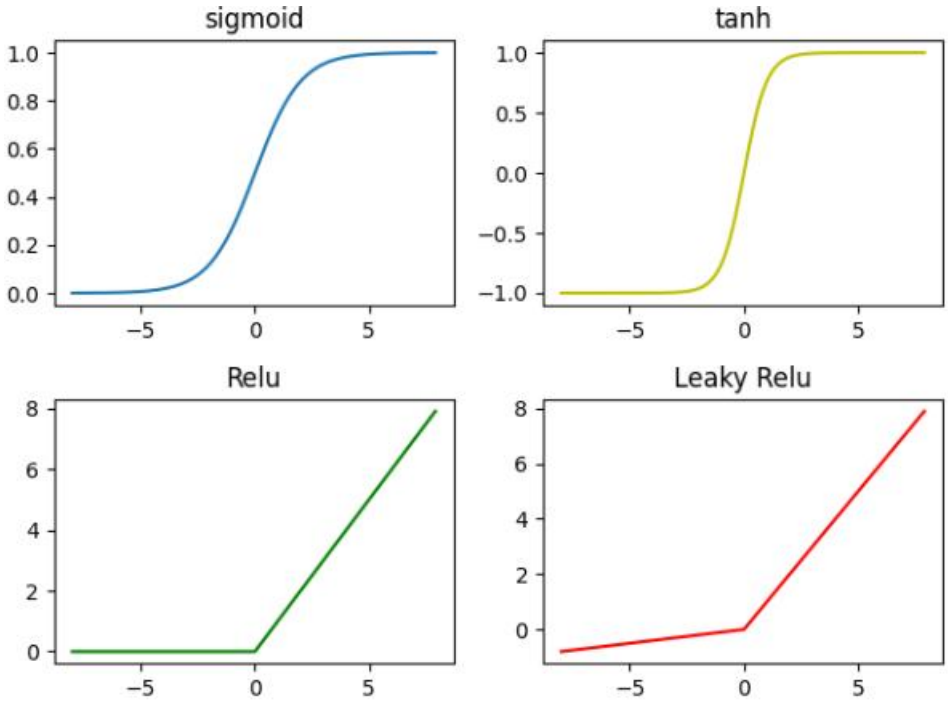

图 2.4 神经网络常用激活函数

在穿越过所有的隐藏层后，会进入到输出层，得到输出值：

$$y = \sigma(z)\theta + b_1$$

从输入层到输出层的过程属于神经网络的前向传播，在这之后需要通过反向传播算法来改变参数值。反向传播算法包括两个要素：损失函数和梯度下降。对于分类神经网络来说，常用的算是函数为交叉熵损失函数：

$$L(y_h, y_r) = -\sum_{i}^{m} [y_r^{[i]} * \log(y_h^{[i]}) + (1 - y_r^{[i]}) * \log(1 - y_h^{[i]})]$$

其中$y_r^{[i]}$表示第 i 个样本的真实值，$y_h^{[i]}$表示第 i 个样本的预测值，也就是神经网络最后一层的输出，如果预测值和真实值存在很大的差异，损失函数的值就会很大，相反，当输出值和真实值非常接近时，损失函数的值就会很小。将所有样本对应的损失之后，进行相加，即可得到对应的总体损失，然后就进入到了优化环节。在神经网络模型中，最常用的优化算法是梯度下降，梯度就是导数，通过对损失函数进行求导，可以得到神经网络中所有参数应该改变的方向和大小。

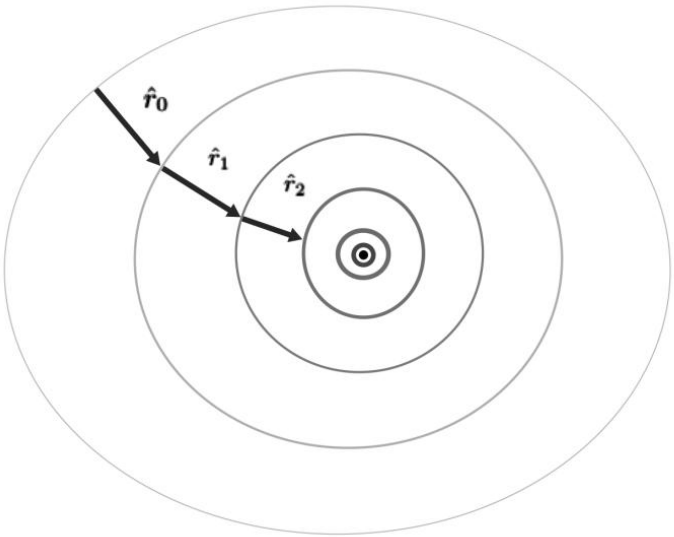

图 2.5 梯度下降原理

将损失函数对神经网络的参数进行反向求导可得：

$$\frac{\partial L}{\partial w^{[i]}} = \frac{1}{m}\frac{\partial L}{\partial z^{[i]}}(A^{[i-1]})^{T}$$

$$\frac{\partial L}{\partial b^{[i]}} = \frac{1}{m}\sum_{i=1}^{m}\left(\frac{\partial L}{\partial z^{[i]}}\right)$$

$$\frac{\partial L}{\partial A^{[I-1]}} = (w^{[i]})^{T}\frac{\partial T}{\partial Z^{[i]}}$$

神经网络参数更新：

$$\mathrm{w}^{[\mathrm{i}]} = \mathrm{w}^{[\mathrm{i}]} - \mathrm{r}\frac{\mathrm{d}\varphi}{\mathrm{dw}^{[\mathrm{i}]}}$$

$$b^{[i]} = b^{[i]} - r\frac{d\varphi}{db^{[i]}}$$

其中 r 表示学习速率，r 越大，参数更新越快，但是精度越低，r 越小速度越慢，但是最终得到的模型精度较高，一般会使用动态 r 使得在刚开始时训练的速度较快，在训练的过程中逐渐减小 r 的大小，使得精度提升。

以上是一轮训练过程，通过迭代，每次分类的误差会下降，当模型达到一定精度以后停止训练，即可得到最终的神经网络模型。在神经网络中存在一些需要调整的超参数，包括神经网络的层数，在上面的介绍中我们仅构造了一个隐藏层，而在实际运用当中我们会构造多个隐藏层，层数的增加有助于模型对样本的深层次特征进行挖掘，此外还有每层神经元的数量，一般来说神经元数量越多，模型

所能够拟合的关系就越复杂。不管是网络的层数还是神经元的数量，并不是越多越好，如果这些数量过多虽然会在训练集上取得非常好的拟合效果，但是在实际的运用当中效果很差，即出现了过拟合问题，神经元网络的层数，每层神经元的数量还有学习速率 r 的大小，都需要通过交叉验证来确定。

2.3 文本分析

文本分析是指通过自然语言处理技术来对文本文档、媒体信息、互联网网页等文本格式内容进行信息抽取，从而产出大量有有价值且能够被各种模型所使用的数据信息。将文本信息量化的方式通常有三种：

（1）Bag-of-words 模型

Bag-of-words 也叫词袋模型，由 Harris et al.(1954)年提出。词袋模型通过将文本当中的每个单词进行 one-hot 编码，然后根据每个单次在文中出现的次数构造一个文本矩阵，这个矩阵就是文本的数量化表示。词袋模型的优点是简单易于理解，但是它假定每个词语之间是独立的没有任何关系，这阻碍了文本信息的理解，而且导致了特征的稀疏性。

（2）word2vec 模型

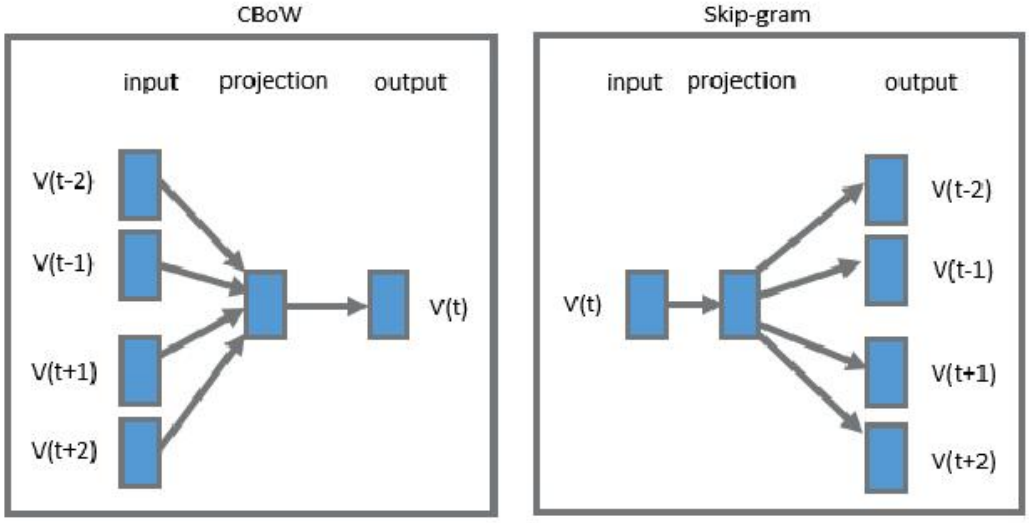

图2.6 word2vec训练模型

word2vec 模型由 Mikolov et al.(2013)提出，这是一种利用神经网络训练词向量的方式，可以将词语转换成维度较低的向量，而且这个向量会保持词语的本来

含义。Word2vec 的本质是将每个单词转换为空间中的一个点，而这个点所在的位置就表示该词语在语义空间中的位置，两个点越接近，其含义也就越接近。如图 2.6 所示，实现 word2vec 的方式有两种，第一种是 CBOW（continuous bag of words），CBOW 利用上下文的单次来对某个中心词进行预测，最后利用梯度下降进行迭代来缩小每次的预测误差，最终从隐藏层的权重矩阵当中获得每个单词的向量表示。第二种是 skim-gram，该方法和 CBOW 正好相反，它是通过中心词来预测上下文某个单次出现的概率来训练模型，最终获得所有单词的向量表示。在所有的词向量训练完之后通过将文本的词向量合并求其均值即可获得某段文本的向量表示。

从 word2vec 的训练过程中可以看出，word2vec 只能表征出每个词语的向量，但是忽略了句子中每个词语的顺序，虽然可以通过求均值来得到每条文本的向量表示，但是表达程度并不一定精确。

（3）doc2vec 模型

针对 word2vec 不能加入单词顺序这一重要信息的缺点，Mikolov et al.(2014)提出了能够结合单词顺序的 doc2vec 模型。

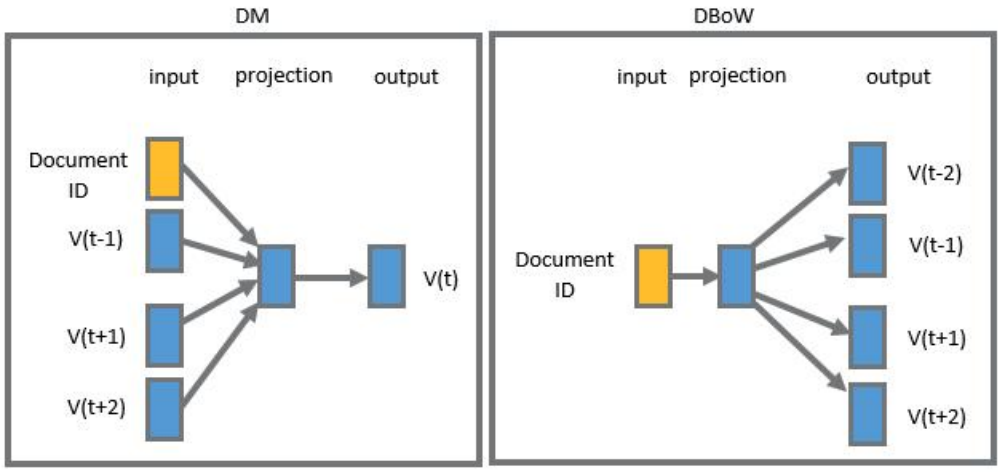

图 2.7 doc2vec 训练模型

如图 2.7 所示，与 word2vec 类似 doc2vec 也可以分为两种形式，分别是 DM 模型和和 DBOW 模型。DM 模型在利用文本数据训练时，会先将每个文档的 ID 和语料库中的所有单词设置成一个 k 维的向量，之后将事先设定的文档向量和其中的上下文单词向量共同输入到模型当中，在一个文档的训练过程中，文档 ID 保

持不变，所有单词共享着同一个文档向量，相当于在预测单词的概率时，都利用了整个句子的语义。DBOW 和 DM 模型的区别在于 DBOW 模型的输入是文档的向量，预测的是该文档中随机抽样的词。

鉴于新闻文本由句子组成，单词的顺序对于整段文本的理解有着一定的价值，所以本文采用 doc2vec 模型作为文本量化的工具。

3 数据收集与处理

3.1 网络爬虫

网络爬虫是指计算机按照用户事先设定的程序或者是脚本来自动地对网页上的某些特定数据进行采集。网络爬虫可以模仿浏览器来自动地进行网页浏览，自动大批量地对用户所需要的网络资源比如文本信息、图片、音频甚至是视频信息进行爬取。Python 是现在使用最广的爬虫编程语言之一，这是因为它提供了大量的第三方库来帮助用户书写爬虫程序。其中第三方库 Scrapy 可以帮助用户制定整个爬虫框架，使用 Scrapy 库首先需要对目标 URL 发送请求，然后等待该网址的响应，在网页做出响应以后即可获得网页上的数据。由于网页上的数据过于杂乱，为了帮助用户从众多杂乱的数据中取得真正有用的信息，python 提供了专门用于解析网页的 BeautifulSoup 模块，该模块可以方便地调用多种解析引擎来对 HTML 网页进行分析。BeautifulSoup 将 HTML 中的信息用树形结构进行保存，这样使用者可以利用平行遍历、上行遍历和下行遍历来解析网页中的相关信息从而获得自己想要的内容。通过在 Scrapy 中对 BeautifulSoup 中的解析功能进行加载就可以获得一个完整的信息爬取和分析的爬虫系统。在提取出我们所需要的信息后可以根据所提取的信息内容进行相应的文件存储，本文仅需要获取网页中的某些文本信息，所以只需要在解析完成后将文本文件存储进表格当中即可。

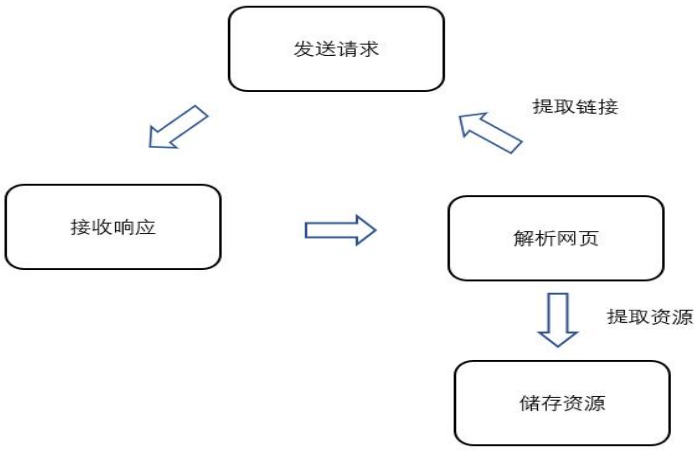

图 3.1 网络爬虫原理

3.2 数据来源

3.2.1 文本数据来源

表 3.1 infobank 中国经济新闻库数据源

新闻源类型	新闻源名称
综合性新闻源	《光明日报》、《人民日报》、《新华日报》、《新民晚报》、《环球时报》、《南方都市报》、《杨子晚报》
专业性新闻源	《中国证券报》、《经济日报》、《上海证券报》、《经济参考报》、《证券日报》、《财联社》、《第一财经》
自媒体新闻源	东方财富、同花顺、雪球财经、和讯网、新浪财经、中金在线

本文使用的文本数据来自于 infobank 数据库，infobank 收集了自 1992 年以来的大量新闻数据，经过二十多年的积累，infobank 已经成为了全球最大的中文信息库之一。为了让所用的数据尽量和金融保持相关，本文只采用了 infobank 中的中国经济新闻库。中国经济新闻库会对每天在中国社会上流转的经济新闻进行收集，如表 3.1 所示，这些新闻的来源主要包括了三个种类。第一类是综合性较强的新闻报纸，比如《人民日报》和《光明日报》，这类报纸经常会报道一些和国家经济形势，方针政策甚至是国际环境有关的新闻，受众广泛，对整个金融市场有着重要的影响，经济新闻库收录了这类报纸中和经济有关的新闻报道。第二类是专业性很强的财经类新闻，比如《中国证券报》和《经济日报》，这类新闻报纸上经常会刊登一些和金融市场有关的新闻，在投资者中有很大的影响力。第三类是一些主流的财经类网站，典型的代表包括东方财富、同花顺、新浪财经等，随着我国移动互联网的不断发展，越多越多的投资者开始通过这类财经网站了解和金融市场有关的新闻，这条渠道种的新闻对投资者的影响也越来越

大，因此有必要多这些网站上的新闻进行收集。这三条渠道基本涵盖了我国经济新闻的主要来源。在时间范围上，为了和股票市场的数据进行对应，只选取了2005 年 4 月 7 日至 2021 年 12 月 2 日的数据。

3.2.2 股指数据来源

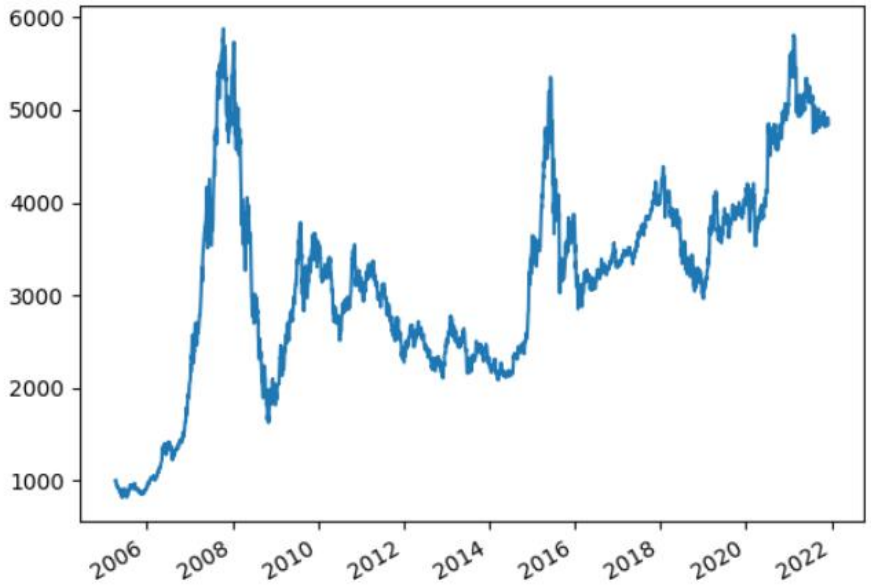

图3.2 沪深300指数历史表现

由于选用的新闻源通常针对的不是某一支股票，而是整个金融市场，所以很难相信从新闻中提取的信息可以预测某只股票的走势。因此，本文将预测标的集中在能够较为准确地反映整个金融市场的投资产品——沪深 300 指数。沪深 300 指数是由上海证券交易所和深证证券交易所与 2005 年 4 月 8 日联合发布的一种能够反映中国 A 股金融运行状况的金融指标，经常被用作投资业绩的评价标准。指数数据来自于 joinquant 的 python API 接口，时间范围是 2005 年 4 月 8 日至 2021 年 12 月 3 日。

3.3 数据预处理

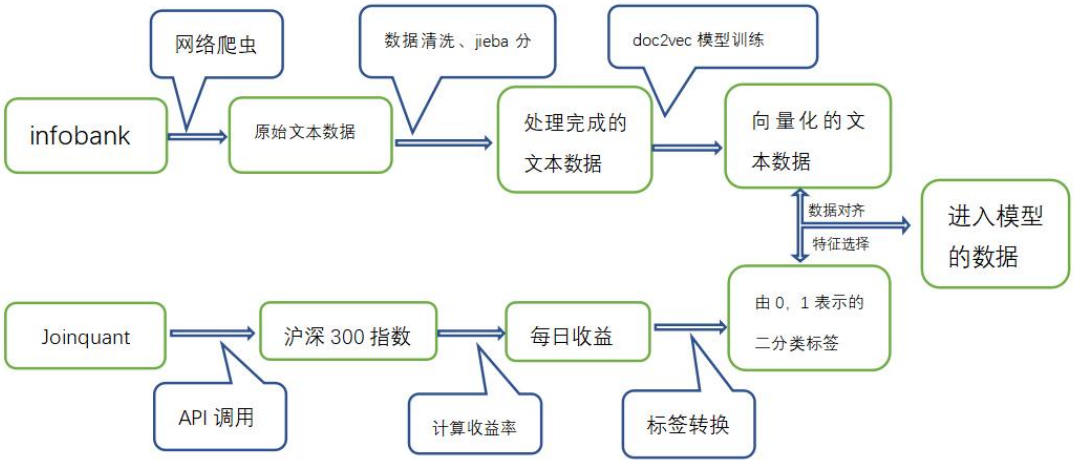

图3.3 数据处理流程

在每条文本数据进入 doc2vec 进行训练之前，必须先将文本切分成单词形式，由于语言的结构形式不一样，中文的分词不能像英文那样直接按单词进行切割，必须要借助一定的算法来帮助分词。现在市场上常见的分词工具有：斯坦福分词器、哈工大的 LTP 分词器和 jieba 分词器，本文选择使用 Python 的第三方模块 jieba 来对新闻文本中的句子进行词语分割。jieba 分词器是通过前缀词典来实现高效的词图扫描，然后检查由句子当中所有可能出现的成词情况，之后将这些可能的单词表示为的一种有向的无环图，最后利用动态规划来分析最大的概率路径，寻找出基于词频最大的切分结果。jieba 分词还允许用户自定义一些想要两外加入的词典，由于本文所涉及到的领域为金融投资领域，所以选取了招商金工团队为金融文本挖掘所构造的招金词酷，然后再通过人工收集的方式对词库进行扩充，最终构造出针对金融投资领域的自定义词典。其中所使用的招金词酷是由 wind 资讯当中的词汇，搜狗金融词库加上招商金工团队所总结出来的专业词汇所构成，这其中包含了大量的证券投资专业术语、金融词汇、财经类词汇等。Jieba 作为一种强大的分词工具一共包含了三种分词模式，分别是精确模式、搜索模式与全模式。其中精确模式指的是让模型对句子实现最精确的切分，把最合适的分词结果作为模型的输出结果。全模式是把句子切分成所有可以成词的词语，提供了所有可能出现的分词结果，但是这种模式不能解决歧义。而搜索引擎模式则是先让句子实现最精确的切割，然后再此基础上对厂词再次进行分割，从而对搜索

引擎的召回率或者是命中率进行提高。本文对三种模式的分词效果进行了对比分析，最后发现搜索引擎模式最终的切分效果最好，其切分出来的词语细粒度更高，所以最后选择了以搜索引擎模式来对新闻文本进行分词。

在进行分词之后，需要对每条文本数据进行清洗，去除当中的数字，标点符号和字母从而减少模型中的噪声，这一部分主要通过正则表达式来完成。在 doc2vec 训练结束之后，可以得到每个单词的向量表示，然后通过对句子中所有单词的向量进行加总求平均，即可得到每个句子的向量表示。

在股票市场的数据处理上，由于本文是通过文本信息来预测未来的涨跌，而不是预测未来收益的具体变动，因此整个任务是属于一个二分类任务。为此，将所有正收益的样本赋值为 1，所有负收益的样本赋值为 0。

在分别处理完文本数据和股指数据以后，还需要数据进行对齐处理。由于股票数据并不是每天都会更新，在节假日时，股票市场会停盘，此时就没有预测标的。而为了能够让文本数据得到充分的利用，将停盘时期所积累的文本信息全都用于下一个交易日的股票走势预测。此外，可能是由于数据库的原因，存在某些天没有相关新闻的情况，虽然这样的情况很少，但是也会阻碍数据的对齐处理，因此根据缺少的新闻，将对应的股票信息删除。经过处理后，新闻条数为 1197493 份，股票收益数为 4016 条。

3.4 特征筛选

因为维度太高，其中并不是所有的信息都和股市有关，如果把这些特征全都加入到模型当中很有可能会出现过拟合现象，因此需要对文本的高维特征进行筛选，本文采用的方法是随机森林算法来进行筛选。

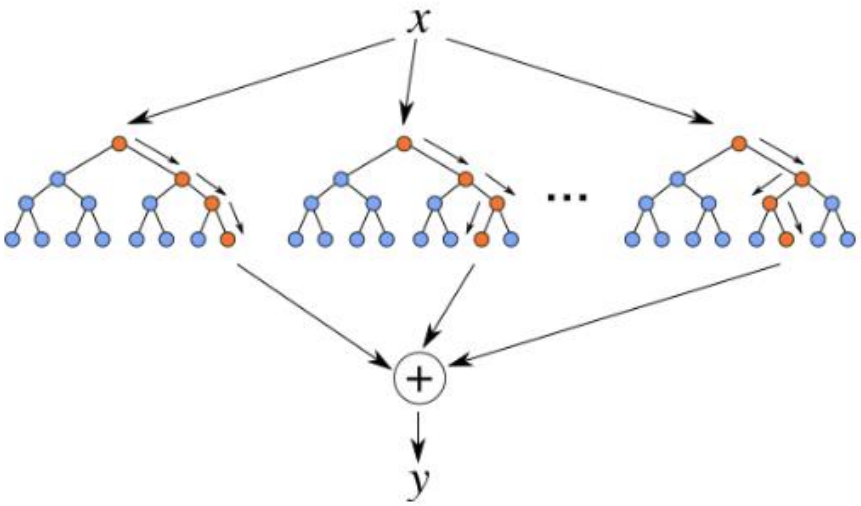

图3.4 随机森林

随机森林是集成学习算法中的一种，该理念最早由何天琴于 1995 年率先提出，之后由 Leo Breiman 和 Adele Cutler 于 2001 年在一篇论文中正式发布。随机森林是对决策树的一种扩展，单棵决策树容易出现过拟合问题，为了解决这个问题，随机森林将 bootstrp aggregating 方法和 random subspace 方法相结合。随机森林和决策树的差异主要体现在两个方面。第一，在样本的选择上不同，决策树由于只有一棵树，所以在训练时会将所有的样本集中起来进行一次完整的训练，而随机森林最终需要生成多颗不同的决策树，因此在对每棵树进行训练时需要获得不同的样本集，在随机森林中，每个子数据集会对原样本进行放回抽样，每棵树在训练时会获得数量相同，但是样本不同的数据集，保证了决策树的多样性。第二个不同之处在于特征的选取，单一的决策树在对样本点进行分割时会考虑到所有的特征，从而找到最佳的分割点，也正因为此，决策树最终形成的边界对样本值非常敏感，鲁棒性相对较差，为此，随机森林会在对每个子数据集进行训练时，会采用随机抽样的方式来选择不同的特征，然后在这些被抽选出来的特征基础上构建各自的决策树，这样有助于决策树在进行分类或者是回归预测时不会过度依赖于某一种特征，而是会综合考虑多种因素，从而提高模型的性能表现。

随即森林的特殊训练过程让其不止能得到一个鲁棒性更好的分类或者是回归模型，还可以对样本的每个特征的重要程度性进行度量。在对森林当中的每棵树进行训练时，我们是对样本进行有放回抽样，因此还有一些样本没有进入到这棵树的训练当中，这部分样本被称为袋外数据（out of bag），将袋外数据放到

决策树模型中进行测试，可以计算出此时模型的预测错误率，这部分错误率又被称为袋外数据误差。接下来，对袋外数据的所有样本特征 X 加入噪声干扰，从而实现对特征的改变，将改变后的数据再次放到模型中进行计算，得到第二个袋外误差，将所有袋外误差的差额相加求平均，该均值越大说明在加入随机误差后，误差率增加的越大，这也说明了这个特征值的改变会对分类结果产生较大影响，从而实现了对特征重要性的度量。

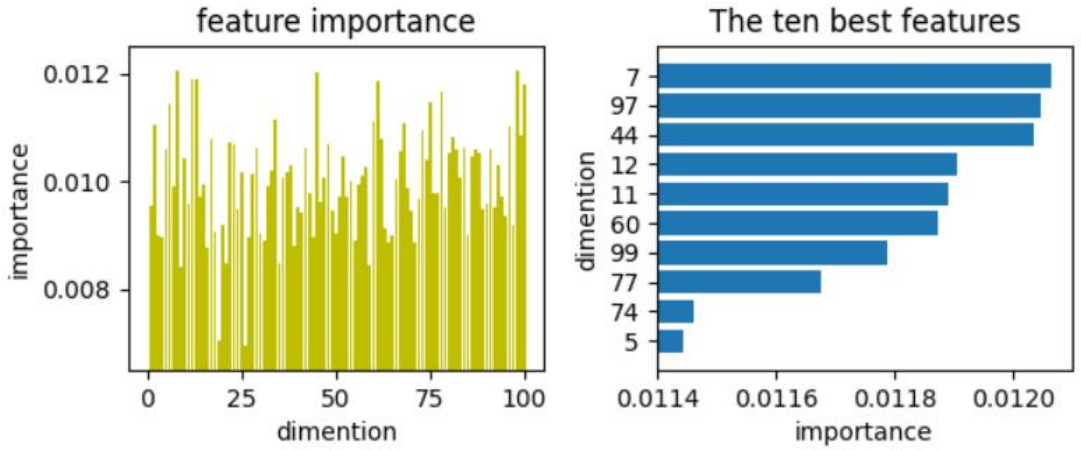

图3.5 随机森林特征筛选结果

图 3.5 展示了利用随机森林计算出的每个文本特征的重要程度，可以看到，并非每个特征的重要程度都相同，因此本文采用最有预测效果的文本特征来当作特征因子，避免维度灾难的出现。

4 基于文本挖掘和机器学习算法的股票预测研究

4.1 训练集与测试集的划分

机器学习的准确性来自与对大量样本的训练，只有训练足够多的样本，机器学习算法才能从中挖掘到正确的规律。另外，为了保证实验结果的有效性，不能将最终要测试的数据放入模型，因此必须要将样本划分成训练集和测试集。本文将 2005 年 4 月 8 日至 2019 年 12 月 2 日的数据定为训练集，将 2019 年 12 月 3 日至 2021 年 12 月 3 日的数据定为测试集，最终得到训练集样本量 3564 条，测试集样本量 487 条。训练集和测试集的沪深 300 指数收盘价如图 4.1 所示：

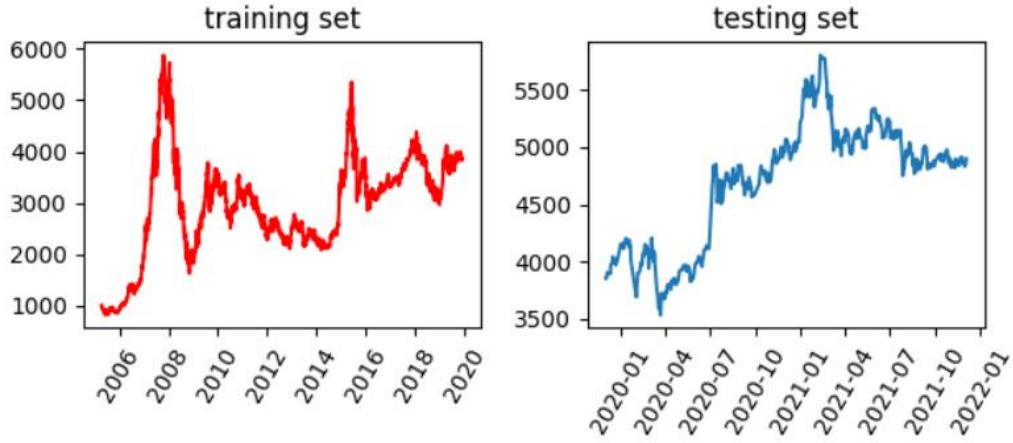

图 4.1 训练集和测试集的划分

4.2 参数调优

在机器学习当中，由于每种算法所涉及到的参数较多，而参数的好坏直接导致了模型的最终表现，机器学习算法中最大的难题不是无法对样本进行非常好的拟合，而是过拟合问题。我们在寻找好的模型时实际上是在寻找合适的参数，好的参数不仅在训练集上表现良好，在测试集上也会出现非常优良的表现。而在机器学习当中，通常会使用交叉验证法来进行参数的调优。交叉验证法的基本思想是将训练集分为两个部分，一个部分是用于模型的训练，通常由于机器学习有非常强的非线性拟合能力，所以训练出来的模型基本上都能在这部分样本上获得非常好的表现，但是这种表现很可能并不是由于模型找到了正确的模式，而仅仅是

因为模型出现了过拟合现象，为此，我们需要第二部分模型没有见到过的样本即验证集来对模型的好坏进行评价测试。如果模型在训练集和测试集上都有着比较好的拟合结果时，说明模型训练的比较好，此时过拟合的风险也相对较低。在训练模型时经常会将交叉验证法和网格搜索等寻参方法一起使用，目的就是为了能够快速找到表现最好的参数。交叉验证法有多种划分，根据训练过程不同可以分为简单的交叉、留一法、留 P 法以及 K 折交叉验证方法。

简单交叉验证法就是简单地将所有训练数据划分为两类，将其中一类作为训练集用于模型的训练，另外一类则留下来用于模型的测试，简单交叉验证只进行以此划分，然后进行模型的训练和测试，这种方式虽然效率很高，但是始终会存在某些样本没有进入到模型的训练当中，实际上造成了对样本的利用率相对较低。留一法则是在对数据集进行切割时，每次只留下一个样本当做模型测试的验证集，剩下的样本都作为模型的训练集，在进行完一次训练后，和测试后，再挑选出另外一个样本作为验证集，其它样本数据作为训练集，重复以上步骤直到每个样本都被充当过一次验证集，这种方法充分利用到了所有的样本信息，训练出来的结果也和整个测试集的期望值最为接近，但是这样的计算需要耗费大量的成本，只适合在样本量较少时使用。留 P 法的运作步骤和留一法基本相同，只不过在选取验证集时是会选择 P（P>1）个样本，与留一法一样，可以对所有数据进行较为充分的利用，但是依旧存在计算量大的问题，不适用数据量较多的情况。对比来看，

这三种验证方法在实际的模型训练当中远没有 K 折交叉验证法的使用频率高。

K 折验证法首先将整个训练集分为 K 个部分，每次进行模型的训练时会将其中的一个部分拿出来作为验证集，其它的部分都用于训练模型，在这一轮结束后会拿出另外一个部分作为验证集，然后其余的部分用于模型的训练，这样重复下去，直到每个部分都被用做一次测试集。在整个训练过程中，每个验证集都会得出自己的实测精度，将所有的误差取其均值，最后即可得到该模型的预测表现，下图展示了当 K 为 10 的交叉验证训练过程。

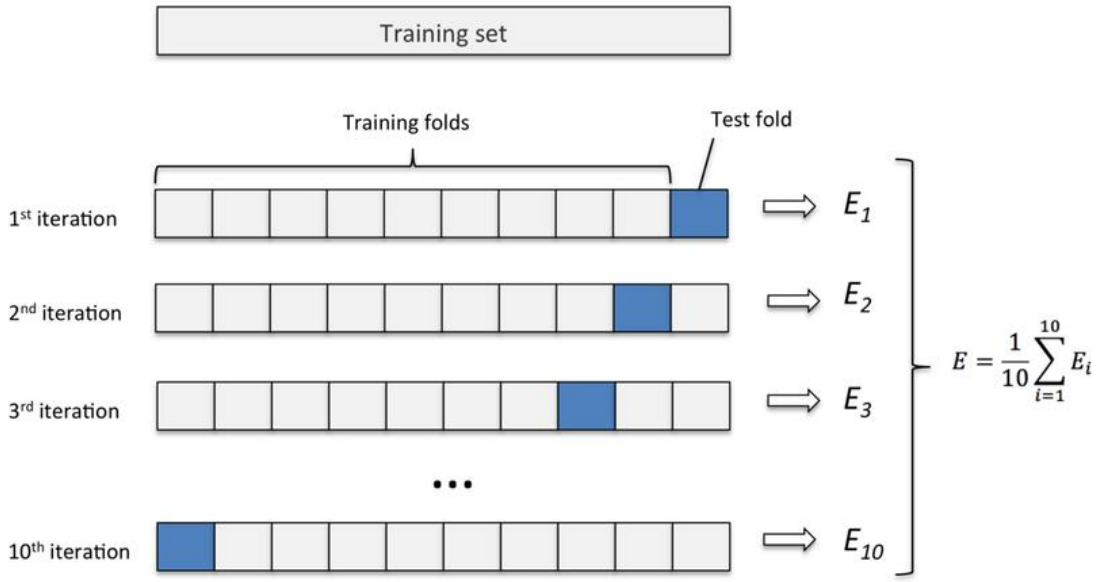

图 4.2　10 折交叉验证

上面提到的集中交叉验证方法有一个共同的假设前提，那就是假设每个样本数据都满足独立同分布的假设前提，即所有样本的取值都不存在相互影响，而且么个样本都取自于同一个分布。但是金融市场上的时间序列可能难以满足这一假设前提，金融时间序列很有可能会受到周期性或者是趋势等不稳定因素的影响，这样其样本数据就很难满足独立同分布的假设条件，所以传统的 K 折交叉验证法并不适用，并且很有可能出现过拟合现象。

时序交叉验证法可以用于处理关于时间序列模型的训练，改交叉验证法并不需要让数据服从独体同分布的假设条件。时序交叉验证法在划分数据集时并不会向 K 折交叉验证那样随机打乱样本的顺序，而是按照时间的先后顺序来一次对训练集和样本进行划分，在划分的过程当中，训练集结束的时间点就是测试集开始的时间节点，而在整个过程当中，都是用过去的历史数据来对模型进行训练，然后使用最新的数据来进行模型的测试，通过这种方式，就避免了将要预测的数据样本划分成了训练集，或者是将要训练的样本用作了验证集。和十折交叉验证法一样，十折时序交叉验证法也需要将整个训练集的样本数据划分为十个部分，然后按照时间的先后顺序将每个部分的数据集打上从第一到第十的标签。在训练时首先将第一个部分的样本数据取出用于模型的训练，然后将第二个部分的样本取出用于模型的测试。在第二轮的训练过程中，用于训练集的样本数据是第一和第

二两个部分，将第三个部分的样本数据用于模型的测试。重复这样的过程，直到最后一个部分的样本数据被用于模型的测试，然后取每轮训练的误差，求其均值作为这个模型的最终表现。和其他的交叉验证法相比，时序交叉验证法只会使用到历史的数据来进行模型的训练，不会使用到未来的数据，所以更适合应用到金融时间时间序列上的模型调优。

4.3 评价指标的选取

针对本文的研究目的，最终选择对模型分类效果和模型所产生的盈利两个方面来进行评价。这是因为股票的上涨和下跌本质上是一个分类问题，所以有必要对模型的分类性能进行评价，另一方面，量化投资模型会根据分类结果来进行投资，我们可以通过投资回测来衡量每种算法最终能够获得的收益。

4.3.1 分类效果评价指标

考虑到股票预测实际上是一个二分类任务，因此有必要对其分类效果进行评价，在分类效果上我们可以通过混淆矩阵来构造评价指标进行评判，混淆矩阵的构造如下表所示：

表 4.1 混淆矩阵

	实际 True	实际 False
预测 True	TT	FT
预测 False	TF	FF

根据混淆矩阵可以构建出准确率、精确率、召回率、F1 分数等四个评价指标。准确率表示所有预测正确的数量占总体数量的比重，精确率表示正确预测为证的占全部预测为证的比例，召回率表示正确预测为正的占全部实际为正的比例,F1 值为精确率和召回率调和均值的两倍，该指标综合考虑精确率和召回率的大小，只有当两者都表现较好时才能取得较好的 F1 分数。

表 4.2 基于混淆矩阵的评价指标

评价指标	公式
准确率	$\text{Accuracy} = \frac{\text{TF} + \text{FF}}{\text{TT} + \text{FF} + \text{TF} + \text{FT}}$
精确率	$\text{Precision} = \frac{\text{TT}}{\text{TT} + \text{FT}}$
召回率	$\text{Recall} = \frac{\text{TT}}{\text{TT} + \text{TF}}$
F1 分数	$\text{F1} = \frac{2}{1/\text{Recall} + 1/\text{Precision}}$

4.3.2 盈利能力评价指标

为了检验最终模型的盈利能力，本文设计了一个模拟投资实验，令初始资金为一万元，每天下午三点收盘之前进行投资决策，在模型给出买入信号时如果此时处于空仓状态，则将所有现金投资进去，如果此时已经持仓则不进行任何操作；在模型给出卖出信号时，如果此时是不是空仓状态，则选择清仓处理，如果此时已经是清仓状态，则不进行任何操作，等待下一次的操作。在回测过程中不考虑交易手续费和交易对市场的冲击。

如表 4.3 所示，为了描述回测结果的好坏，通常会使用累计收益率、年化收益率、最大回撤、夏普比率等指标。

表 4.3 评价指标说明表

指标名称	指标公式
累计收益率	$cum\_profit(T) = \frac{total\_money_T - 1}{1} \times 100\%$
年化收益率	$year\_profit(T) = (1 + cum\_profit)^{\frac{250}{T}} - 1$
最大回撤	$\max\_retrace(T) = \frac{\max(P_x - P_y)}{P_x}$
夏普比率	$sharpe\_ratio = \frac{(R\_p - R\_f)}{\sigma\_p}$

累计年化收益率就是总的投资收益率，它表示在整个投资周期中，最终所获得的收益。年化收益率指的是在整个投资周期当中，每年所获得的收益率是多少，值得注意的是与银行存款、债券等投资标的不同，股票市场的一年的有效投资日一般定为 250 天。最大回撤是指股票从历史最高点到历史最低点的跌幅，该指标主要是对投资标的风险的一种衡量，即使某些标的在很长一段时间后呈现的是盈利状态，但是在之前的投资周期中，股票呈现了很大幅度的下跌，一些风险承受能力较弱，或者是债务比例过高的投资者会因为跌幅太大而过早地退出股票市场，承认损失，所以计算最大回撤在后续的投资当中有很大的参考价值。夏普比率由经济学家 William Sharpe 提出，夏普比率又被称作是夏普指数，该指标也是用于衡量金融资产过去业绩表现。夏普比率指的是投资者每承担一单位风险时能够产生多少的超额报酬，是一个能够综合反映投资风险和投资收益的指标，如果夏普比率为正值则表示此投资标的的报酬高于波动风险，如果夏普比率为负值，则表示此时的收益无法覆盖波动风险，如果夏普比率的值越大，说明该投资方案的表现越佳。

4.4 未加入文本特征的回测结果分析及评价

表 4.4 未加入文本特征的分类结果

	准确率	精确率	召回率	F1 分数
支持向量机	0.4913	0.5123	0.5217	0.5169
XGBoost	0.5014	0.5245	0.5397	0.5320
神经网络	0.5167	0.5190	0.5217	0.5203

表 4.4 展示了在未添加文本特征而仅使用最基本的开盘价、收盘价、成交量、累计成交额等数据进行模型训练，由于这些都是市场上最常见的数据，市场上的很多投资者会用这些数据训练模型，在市场的激烈竞争之下，这些数据很难为模型带来能够获取超额收益的特征，无论算法多么强大，在特征没有足够信息的情况下，也很难预测股市的波动。

表 4.5 未加入文本特征的收益回测

	累计收益	年化收益	最大回撤	夏普比率
Buy & hold	27.26%	13.17%	18.19%	0.58
支持向量机	5.68%	2.88%	17.55%	0.46
XGBoost	20.10%	9.86%	14.27%	0.63
神经网络	11.86%	9.15%	19.24%	0.56

表 4.5 展示了在未加入文本特征时，各个模型投资回测时的收益表现，可以看出由于缺乏有用的特征，各个模型的表现在大部分指标上表现都不如买入并持有沪深 300 指数。

4.5 加入文本特征后回测结果分析及评价

表 4.6 加入文本特征的分类结果

	准确率	精确率	召回率	F1 分数
支持向量机	0.5425	0.5321	0.5533	0.5425
XGBoost	0.6392	0.6015	0.5925	0.5970
神经网络	0.5329	0.5536	0.5849	0.5688

在加入文本特征后，支持向量机、XGBboost、神经网络在分类任务中的性能有了明显的提升，这意味文本特征中有着能够预测股市的重要信息，由于使用的该特征还未被广泛使用，所以能够为使用这些特征的投资者带来一定程度的超额收益。

在各个机器学习算法的表现上，XGBoost 在准确率、精确率、召回率上的表现分别为 0.6392，0.6015，0.5925 在所有指标上的提升最为明显，分类效果做好。从三个机器学习模型背后的原理和金融市场中的交运作情况进行分析，可得出以下两种原因：第一，XGBoost 属于一种集成学习模型，这种模型的学习过程是通过多个弱学习器进组合，相比单一模型能够缓和过拟合问题，因此在回测时能够获得较高的正确率；另外，XGBoost 相比于支持向量机和普通神经网络出现时间

更短，算法更为新颖，此时市场上使用该算法的投资者或者是机构还相对较少，这也就意味着实际运用中模型的同质性相对较低，因而能够取得比其它模型效果更好的表现。

表 4.7 加入文本特征的收益回测

	累计收益	年化收益	最大回撤	夏普比率
Buy & hold	27.26%	13.17%	18.19%	0.58
支持向量机	22.08%	10.78%	9.44%	0.7
XGBoost	45.88%	21.39%	14.64%	1.55
神经网络	30.01%	14.42%	12.86%	0.95

表 4.7 展示了加入了文本特征后各个模型的回测表现，可以看到，在文本特征加入后，回测的结果有了很大的提升，其中 XGBoost 模型的年化收益高达 21.39%，夏普比率也达到了 1.55，在各项指标上都要好于沪深 300 指数本身的表现。另外，神经网络的各项指标也均好于简单的买入持有策略。支持向量机在累计收益上不如原指数，但是由于其在最大回撤上的表现要好于原指数，整体下来还是有着更好的夏普比率。

5 总结与建议

5.1 研究总结

本文将文本挖掘技术和机器学习技术应用到金融投资领域，将股票市场的预测问题当作是一个二分类问题，然后提出了一整套利用文本数据和机器学习算法来进行量化投资的方法。具体来说，本文的所作的工作与最终得到的结论可以概括为以下几个方面：

（1）通过 doc2vec 自然语言处理技术，将大量的财经新闻文本转化为可以被其各种机器学习算法所使用的结构化数据，与以往的 bag of words 技术不同，本文所使用的技术可以更加准确地提炼出文本中的相关信息。

（2）考虑到金融市场中可以使用的数据量较少，而文本特征的维度过高，如果直接使用所有的特征进行训练，很容易产生过拟合问题，影响模型的精度。因此在获取文本特征以后，对其使用随机森林算法从中筛选出重要程度最高的十个文本特征，将其用于模型的训练。

（3）引入了支持向量机（SVM）、XGBoost 模型、神经网络三种不同类型的机器学习算法，分别在没有使用文本特征和使用文本特征上的数据集上进行训练和测试，结果表明文本特征的加入可以提高模型的预测效果，而其中 XGBoost 的表现最为优异。

5.2 启示与建议

（1）本文的研究对我国的金融监管有一定的启示，由于新闻媒体的内容对于投资者的行为有着非常重要的影响，所以政府应当加大对于新闻媒体的监督。不合理的言论容易引发市场的异常，特别需要防范一些投机分子利用媒体手段来操纵市场信息，进而实现非法获利的目的。而在市场发生异常波动的时候，政府还可以通过各大新闻媒体实现“议程设置”，引导投资者回归理性。

（2）本文的研究对于推进我国金融领域人工智能战略有一定的启示。随着数据的积累和计算机算力的提高，人工智能在各行各业的重要性不断提高，毫无疑问，作为一项通用性技术，人工智能技术可以在金融行业的各个领域发挥重要

作用，能够有效地减少成本和降低风险。本文的研究主要是通过研究文本挖掘技术和各种机器学习算法来辅助金融投资公司的发展，这是是人工智能在金融投资领域的一个典型运用。人工智能技术在金融领域的应用远不止于此，通过合理的政策引导，人工智能技术很有促进我国金融行业的进一步发展。

现阶段我国的金融机构，特别是传统金融机构积累了大量的数据，这些数据包括传统的结构化的数据，也有大量传统模型无法处理的非结构化数据，在传统的金融管理当中，金融机构大多使用的数据都是标准的结构化数据，而由于技术的限制，大量非结构的数据比如视频、文本、电话录音等信息未能发挥重要的价值，这些非结构化的数据保存在金融机构的数据库中仅仅是让其作为一种备案方式，为金融业务的开展留下业务凭证，在没有业务需求时一般用不到这些数据，而且在对这些数据进行处理时，可能需要大量的人力参与，增加了整个行业的成本。在人工智能时代，政府应当引导金融机构往智能化方向转型，一方面可以让金融机构利用计算机，机器学习算法来处理以往需要人工参与的项目，减少整个行业的运作成本。另一方面，可以引导金融机构使用人工智能技术对长久累积的非结构化数据进行挖掘，从而辅助金融机构的经营决策。

值得注意的是，虽然我国的金融行业有了大量数据的积累，但是依旧存在数据形式多样，存储方式不同，存储结构各异等问题，这不利于机器学习算法的使用，为此，国家应当对各个金融机构的数据存储方式提出规范，除此之外，还应该打破数据壁垒，让数据实现互联互通，这样才能让人工智能技术在金融领域中更好地落地。

（3）本文所得出的结论对我国的资产管理行业的实践发展有着重要的启示。人工智能在各个领域都起到了越来越重要的作用，但是对于资产管理这个行业，还在很大程度上依靠人的经验，智能化程度还相对较低。本文在经过研究后发现，利用自然语言处理技术来让计算机通过文本信息从而进行自动化的投资能够切实有效地提升资产投资的业绩表现，这为我国的量化投资的实践提供了一个有效的新工具。在进行实际应用时，金融投资公司可以使用更多的资源来对本文的研究进行优化改进，这主要包括三个方面。第一，通过加大文本的数据量来让模型得到更多的训练，自然语言处理最重要的步骤就是收集尽可能多的数据来让模型从中学习到相应的知识，本文由于所能接触到的资源有限，最终训练得到的模型

效果还存在改进空间，资产管理公司可以通过人力、自动化爬虫或者是市场化交易等手段收集到更多的文本数据，为模型的训练提供坚实的数据基础。第二，引进更加先进的自然语言处理模型，与其它传统的机器学习算法不同，神经网络有着更高的扩展性，有着高度的可自定义性，也正是由于神经网络的出现，人工智能在近年来的发展十分迅速，自然语言处理模型也在不断迭代升级，其中以 BERT、ELMO、GPT 等为代表的预训练语言处理模型在多项 NLP 任务上都取得了不错的性能提升，受到了各界的广泛关注，这些模型通常上体量非常大，对这些模型进行训练和使用需要得到更大的算力支持，但是通常会取得更好的结果，资产管理公司应当加大这方面的投入。第三，将自然语言模型提取的文本信息和传统的技术性指标与基本面指标相结合，从而寻找更加可靠的投资方案。传统的技术性指标可以反映出当前市场的价格变动情况，而基本面指标可以反映出公司、行业、国家层面的运行情况，在这两种指标的基础上添加从文本数据中挖掘来的信息，可以使得投资模型对整个市场上的信息有着更加充分地理解从而提高投资模型的效率。

（4）本文对人工智能技术在金融投资研究当中的运用具有重要的启示。人工智能技术在金融投资领域的作用主要有两个方面：一是利用自然语言处理技术从文本中提取文本信息或者是利用图像识别技术来快速地从图像甚至是视频当中获取有效信息，这些新型变量的提取为金融投资的研究提供了新的原材料；第二是在金融投资领域上利用各种机器学习算法提升其预测能力，由于金融领域的复杂性，很多变量之间的关系很难满足线性假设，因此难以使用线性模型进行量化处理，而机器学习方法可以高效地挖掘出各个变量之间存在的非线性关系，进而提高模型的预测能力。本文将自然语言处理技术和机器学习算法相结合，证明了该方法的有效性。未来的学者可以进一步研究人工智能技术在金融投资领域中的运用。随着互联网数据量的不断累积和深度学习算法的不断发展，我们已经可以让计算机帮我们从非结构化的文本数据中获取各种隐藏的量化因子，这些新的因子的出现将帮助学者们对复杂的金融现象进行更深入的研究。而各种机器学习算法的出现，为金融领域的研究特别是金融资产价格的研究提供了强有力的工具，本文将三种机器学习算法用于股票市场价格的变动为该方面的研究提供了一个典型范例，以启发其它金融学者的进一步研究。

（5）本文的研究对个人投资者有一定的启示。随着互联网的不断发展，投资者可以越来越方便地接触到各种媒体新闻，而这些新闻也在潜移默化地影响着投资者的决策行为，但是投资者应当注意，这些被网络媒体大肆报道的新闻很有可能是片面的，如果仅仅因为媒体的报道而进行买卖，很有可能会陷入亏损。因此，投资者在看到琳琅满目的各种新闻时应当理性对待，总结经验，学会如何分辨信息的真实性和重要性，提高防范意识，面对媒体的炒作，做到有自己的投资逻辑而不是盲目跟风，最终做出合理的投资决策。

参考文献

[1]Blaufus, Kay, Axel Möhlmann, and Alexander N. Schwäbe. "Stock price reactions to news about corporate tax avoidance and evasion." Journal of Economic Psychology 72 (2019): 278-292.

[2]Burggraf, Tobias, Ralf Fendel, and Toan Luu Duc Huynh. "Political news and stock prices: evidence from Trump's trade war." Applied Economics Letters 27.18 (2020): 1485-1488.

[3]Behera, Ranjan Kumar, et al. "Co-LSTM: Convolutional LSTM model for sentiment analysis in social big data." Information Processing & Management 58.1 (2021): 102435.

[4]Carvalho, Jonnathan, and Alexandre Plastino. "On the evaluation and combination of state-of-the-art features in Twitter sentiment analysis." Artificial Intelligence Review 54.3 (2021): 1887-1936.

[5]Carta, Salvatore, et al. "Multi-DQN: An ensemble of Deep Q-learning agents for stock market forecasting." Expert systems with applications 164 (2021): 113820.

[6]Chauhan, Priyavrat, Nonita Sharma, and Geeta Sikka. "The emergence of social media data and sentiment analysis in election prediction." Journal of Ambient Intelligence and Humanized Computing 12.2 (2021): 2601-2627.

[7]Harris, Zellig S. "Distributional structure." Word 10.2-3 (1954): 146-162.

[8]Hah D W, Kim Y M, Ahn J J. A study on KOSPI 200 direction forecasting using XGBoost model[J]. The Korean Data & Information Science Society, 2019, 30(3): 655-669.

[9]Javed Awan, Mazhar, et al. "Social media and stock market prediction: a big data approach." MJ Awan, M. Shafry, H. Nobanee, A. Munawar, A. Yasin et al.," Social media and stock market prediction: a big data approach," Computers, Materials & Continua 67.2 (2021): 2569-2583.

[10]Mittermayer, M-A. "Forecasting intraday stock price trends with text mining techniques." 37th Annual Hawaii International Conference on System Sciences, 2004. Proceedings of the. IEEE, 2004.

[11]Mikolov, Tomas, et al. "Efficient estimation of word representations in vector space." arXiv preprint arXiv:1301.3781 (2013).

[12]Mikolov, Tomas, et al. "Distributed representations of words and phrases and their compositionality." Advances in neural information processing systems. 2013.

[13]Mittermayer, M-A. "Forecasting intraday stock price trends with text mining techniques." 37th Annual Hawaii International Conference on System Sciences, 2004. Proceedings of the. IEEE, 2004.

[14]Nikou M, Mansourfar G, Bagherzadeh J. Stock price prediction using DEEP learning algorithm and its comparison with machine learning algorithms[J]. Intelligent Systems in Accounting, Finance and Management, 2019, 26(4): 164-174.

[15]Poornima A, Priya K S. A comparative sentiment analysis of sentence embedding using machine learning techniques[C]//2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS). IEEE, 2020: 493-496.

[16]Ramadhani A M, Goo H S. Twitter sentiment analysis using deep learning methods[C]//2017 7th International annual engineering seminar (InAES). IEEE, 2017: 1-4.

[17]Salisu, Afees A., and Xuan Vinh Vo. "Predicting stock returns in the presence of COVID-19 pandemic: The role of health news." International Review of Financial Analysis 71 (2020): 101546.

[18]Salton, Gerard, and Michael J. McGill. Introduction to modern information retrieval. mcgraw-hill, 1983.

[19]Sun A, Lachanski M, Fabozzi F J. Trade the tweet: Social media text mining and sparse matrix factorization for stock market prediction[J]. International Review of Financial Analysis, 2016, 48: 272-281.

[20]Samarawickrama A J P, Fernando T G I. A recurrent neural network approach in predicting daily stock prices an application to the Sri Lankan stock market[C]//2017 IEEE International Conference on Industrial and Information Systems (ICIIS). IEEE, 2017: 1-6.

[21]Terra Vieira, Samuel, et al. "Q-meter: Quality monitoring system for telecommunication services based on sentiment analysis using deep learning." Sensors 21.5 (2021): 1880.

[22]Thavareesan S, Mahesan S. Sentiment analysis in tamil texts: A study on machine learning techniques and feature representation[C]//2019 14th Conference on Industrial and Information Systems (ICIIS). IEEE, 2019: 320-325.

[23]Usmani M, Adil S H, Raza K, et al. Stock market prediction using machine learning techniques[C]//2016 3rd international conference on computer and information sciences (ICCOINS). IEEE, 2016: 322-327.

[24]Yoshihara A, Fujikawa K, Seki K, et al. Predicting stock market trends by recurrent deep neural networks[C]//Pacific rim international conference on artificial intelligence. Springer, Cham, 2014: 759-769.

[25]部慧,解峥,李佳鸿,吴俊杰.基于股评的投资者情绪对股票市场的影响[J].管理科学学报,2018,21(04):86-101.

[26]段江娇,刘红忠,曾剑平.中国股票网络论坛的信息含量分析[J].金融研究,2017(10):178-192.

[27]贺康,宋冰洁,刘巍.年报文本信息复杂性与资产误定价——基于文本分析的实证研究[J].财经论丛,2020(09):64-73.

[28]姜富伟,孟令超,唐国豪.媒体文本情绪与股票回报预测[J].经济学(季刊),2021,21(04):1323-1344.DOI:10.13821/j.cnki.ceq.2021.04.10.

[29]姜富伟,马甜,张宏伟.高风险低收益？基于机器学习的动态 CAPM 模型解释[J].管理科学学报,2021,24(01):109-126.

[30]刘海飞,许金涛.互联网异质性财经新闻对股市的影响——来自中国互联网数据与上市公司的证据[J].产业经济研究,2017(01):76-88.

[31]陆静,周媛.投资者情绪对股价的影响——基于 AH 股交叉上市股票的实证分析[J].中国管理科学,2015,23(11):21-28.

[32]龙文,毛元丰,管利静,崔凌逍.财经新闻的话题会影响股票收益率吗?——基于行业板块的研究[J].管理评论,2019,31(05):18-27.DOI:10.14120/j.cnki.cn11-5057/f.2019.05.002.

[33]吕华揆,刘政昊,钱宇星,洪旭东.异质性财经新闻与股市关系研究[J].数据分析与知识发现,2021,5(01):99-111.

[34]牛枫,叶勇,陈效东.媒体报道与 IPO 公司股票发行定价研究——来自深圳中小板上市公司的经验证据[J].管理评论,2017,29(11):50-61.DOI:10.14120/j.cnki.cn11-5057/f.2017.11.005.

[35]苏治,卢曼,李德轩.深度学习的金融实证应用:动态、贡献与展望[J].金融研究,2017(05):111-126.

[36]苏治,傅晓媛.核主成分遗传算法与 SVR 选股模型改进[J].统计研究,2013,30(05):54-62.

[37]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.

[38]王淑燕,曹正凤,陈铭芷.随机森林在量化选股中的应用研究[J].运筹与管理,2016,25(03):163-168+177.

[39]徐巍,陈冬华.自媒体披露的信息作用——来自新浪微博的实证证据[J].金融研究,2016(03):157-173.

[40]杨晓兰,沈翰彬,祝宇.本地偏好、投资者情绪与股票收益率:来自网络论坛的经验证据[J].金融研究,2016(12):143-158.

[41]游家兴,吴静.沉默的螺旋:媒体情绪与资产误定价[J].经济研究,2012,47(07):141-152.

[42]尹海员,吴兴颖.投资者高频情绪对股票日内收益率的预测作用[J].中国工业经济,2019(08):80-98.

[43]杨洁,詹文杰,刘睿智.媒体报道、机构持股与股价波动非同步性[J].管理评论,2016,28(12):30-40.

[44]杨青,王晨蔚.基于深度学习 LSTM 神经网络的全球股票指数预测研究[J].统计研究,2019,36(03):65-77.

[45]赵琪,徐维军,季昱丞,刘桂芳,张卫国.机器学习在金融资产价格预测和配置中的应用研究述评[J].管理学报,2020,17(11):1716-1728.

[46]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.

[47]张倩玉,严冬梅,韩佳彤.结合深度学习和分解算法的股票价格预测研究[J].计算机工程与应用,2021,57(05):56-64.

后 记

时光飞梭，转眼间，我在兰州财经大学读研的时光就要结束了，在兰财的三年时间，我受到了老师的悉心指导，也到到了舍友和同学的诸多帮助，我相信在兰财的这些年的经历一定会让我受益终身。在毕业答辩之前，我想借此机会表达我对学校、各门学科的老师、辅导我写作的导师、还有共同度过这些年充实时光的同学们的感激之情。

首先，我一定要感谢我的导师，我的导师治学十分严谨，在我的学术道路上给予了我诸多帮助，特别是在我论文遇到瓶颈，需要专业人士的建议时，老师总能从大局上看到我论文问题的所在，然后给予我专业可靠的意见，如果没有导师对我不厌其烦地进行指导，我就不可能完成我的这篇论文。另外我的导师在生活当中十分平易近人，正是因为导师的温柔，我才能放心地说出自己在学术上和生活上所遇到的困惑。同时，我还要感谢开题和预答辩过程中给予我指导的各位老师们，他们犀利的评论让我发现了自己在做研究时从未想到的问题，更为自己的写作拓宽了思路，使得我的论文更加完善。

此外，我还要感谢我的父亲和母亲，正是因为他们勤劳的双手，我才能够在求学的道路上免去后顾之忧，让我可以在兰财之中尽情丰富自身的学识，也正是因为父母的言传身教，我才能够明白生活的意义。在之后的工作中，我将铭记父母对我的指导，去做更多有意义的事情，同时在社会上传递更多的正能量。

学无止境，虽然我在兰财的生活即将结束，但是我对知识的渴望仍未有半分衰弱，以后无论在哪里，我都将铭记兰财各位老师对于我的指导，积极地将自己的所学运用到社会的建设当中，并且时刻督促自己在学习的道路上勇攀高峰。