一、研究背景介绍

股票市场指股票发行、买卖、交易的市场，是证券市场的一部分。由于其与社会经济的密切联系,又被称为经济和金融活动的"晴雨表"。中国股票市场经过三十余年的快速发展,对我国经济社会建设和社会经济发展均产生了深远的社会影响。然而, 由于中国股市的频繁波动, 加上投资者对股价走势的错误判断, 往往会导致市场危机, 严重的还会演变成"股市灾难"。因此, 能否有效预测股票市场以及如何更精准预测股票市场对于学术界和金融界都是广泛关注的课题。

传统的股票预测方法使用统计原理和计量经济学模型来描述金融时间序列数据，比如在计量经济学中，ARCH模型能准确地模拟时间序列变量的波动性的变化，它在金融工程学的实证研究中应用广泛，使人们能更加准确地把握风险（波动性），尤其是应用在风险价值（Value at Risk）理论中，在华尔街是人尽皆知的工具。在统计中，ARIMA模型（自回归模型和移动平均线），即box-Jenkins模型，是主要模式的时间序列。许多学者根据股票历史价格走势的时间序列数据进行回归建模，运用回归模型预测股价的短期变化。如吴玉霞和温欣[1]选取"华泰证券"250期的股票收盘价作为时间序列实证分析数据,通过建立ARIMA模型对创业板市场股票价格变动的规律和趋势进行了预测。实证结果表明, 该模型短期动态、静态预测效果较好, 可以为投资者和企业在进行相关决策时提供有益参考。但传统的回归方法对样本量和分布程度的要求较高，且难以保障预测的精度和稳定性，导致结果缺乏普遍性，因此，适用性十分有限，并不能满足大多数的应用场景。

图表, 条形图

描述已自动生成

资料来源 [5]

近年来, 人工智能应用技术已经逐渐开始迅速发展兴起。深度学习技术作为发展现代科学人工智能的一个重要技术分支, 在图像识别、语音学和视觉图像识别、自然语言处理等诸多应用领域已经成功取得了显著的科学研究应用成就。机器学习算法是一类从数据中自动分析获得规律，并利用规律对未知数据进行预测的算法。常用的机器学习算法包括NN，SVM，ANN等，对于算法的详细内容我们将在后续综述中进行介绍。监督学习从给定的训练数据集中学习出一个函数，当新的数据到来时，可以根据这个函数预测结果。监督学习的训练集要求是包括输入和输出，也可以说是特征和目标，这就使得该算法相较于传统的股票预测方法，有更大的优势，不需要对数据那么高精度的限制。在金融大数据领域, 机器学习算法的应用主要集中在三个方面: 一是利用机器学习来建立或改善投资策略; 二是利用机器学习挖掘金融大数据，获取更多信息; 三是利用机器学习预测金融市场的波动和价格，也是关注度相对最集中的方面，我们综述将更加关注这个方面。

二、金融大数据的特征与内涵

大数据不是简单的"大量数据", 而是一个抽象概念, 它不仅包括历史数据, 还包括每时每刻产生的数据, 甚至包括未来将产生的数据。它具有规模大、功能多、更新快的特点。大数据存在于工业、农业、消费、金融、媒体等各个行业。其中,金融服务大数据以其庞大的应用规模和重要的学术研究应用价值已引起专家学者们的广泛高度关注。通过对其进行实时分析，可以为互联网金融机构提供客户全方位信息，通过分析和挖掘客户的交易和消费信息掌握客户的消费习惯，并准确预测客户行为，使金融机构和金融服务平台在营销和风控方面有的放矢，在竞争中取得更大的优势。许多学者认为, 大数据技术的使用将给金融领域带来深刻的变化, 一群企业积极运用这些科技手段使得金融服务变得更有效率，因而形成经济产业， "金融科技（FinTech）"的概念也随之出现。金融科技可以说是一种新型的解决方案，这种方案对于金融服务业的业务模式、产品、流程、和应用系统的开发来说，具有强烈颠覆性创新的特性。

金融大数据按行业可分为证券大数据、银行大数据和保险大数据。证券大数据包括股票市场、债券市场和衍生品市场的历史市场数据和交易数据。银行大数据包括存款额、货币量和信用卡用户的海量数据等指标。保险大数据包括各种理赔数据、保险数据等。这些金融大数据都来自于现实生活，具有地域跨度大、形式多样、结构不同的特点。在大数据金融时代，大量的金融产品和服务是通过网络呈现的，包括固定网络和移动网络。基于大数据的风险管理理念和工具，在大数据金融时代，风险管理的理念和工具也将有所调整。此外，在大数据金融时代，金融产品和服务的消费者和提供者之间的信息不对称性大大降低。所以，大数据金融无疑是高效的。许多流程和业务都是在线启动和完成的，有些行动是自动化的。正确的产品以正确的方式、在正确的地点、在正确的时间交付给正确的消费者。同时，强大的数据分析能力可以使金融业务非常高效，交易成本将大大降低。与数据收集和整理相比，研究的难度主要在于数据处理和挖掘。我们必须使用各种工具和技术来对我们收集的大量数据进行分类、处理和挖掘。然而，市场环境是不可预测的，很难像处理传统数据一样找到适用性强的传统策略，这也使得新兴的大数据技术显得十分重要。

大数据技术在经济金融领域的一个重要价值在于统计和预测。通过大量数据的一系列计算方法,对一段时间内的趋势进行预测,预测结果更加准确。随着科技的发展,金融大数据的内容开始从不同的角度被填充。学者们利用爬虫技术将文本特征引入模型, 构建投资者情绪和分析方法等指标研究，得到了一些新的结果。如顾文涛等构建了适用于金融投资领域的财经新闻情感词典来对财经新闻进行文本分析,同时构造了新的预测模型:将财经新闻文本中所含的情感量化为情绪指数并与时变密度函数相结合,得到时变加权密度模型。并在此基础上以模型评分为权重组合多个预测模型构建出评分加权模型用于股票收益率预测。结果显示,加入情绪指数能有效提高模型预测能力,而评分加权模型的预测能力则在此基础上更进一步,在准确率以及评分规则上基本达到双重最优。赵明清、吴胜强[4]结合百度指数,利用时差相关系数和随机森林选取微博搜索初始关键词,通过爬虫技术获取微博文本,利用文本挖掘技术对微博文本作分词处理,判断分词后的微博情感倾向,分析影响微博影响力的相关因素,以信息增益确定微博权重，最终得到了两个具有良好预测功能的模型。越来越多的学者将研究视角转向机器学习技术,对金融大数据进行处理和分析。

[1] 吴玉霞,温欣.基于ARIMA模型的短期股票价格预测[J].统计与决策,2016(23):83-86.DOI:10.13546/j.cnki.tjyjc.2016.23.051.

[2] Alt, R., Puschmann, T. (2012). The rise of customer-oriented banking - electronic markets are paving the way for change in the financial industry. Electronic Markets, 22(4), 203-2015.

[3] 顾文涛,王儒,郑肃豪,杨永伟.金融市场收益率方向预测模型研究——基于文本大数据方法[J].统计研究,2020,37(11):68-79.DOI:10.19343/j.cnki.11-1302/c.2020.11.006.

[4] 赵明清，吴圣强. 基于微博情感分析的股市加权预测方法研究[J]. 数据分析与知识发现，2019，3( 2) : 43 - 51.

[5] Deepak, K., Kumar, S.P., & Rajit,V. (2021).A systematic review of stock market prediction using machine learning and statistical techniques[J]. Materials Today: Proceedings