广东外语外贸大学

GUANGDONG UNIVERSITY OF FOREIGN STUDIES

数据挖掘（课程设计）

题　　目：　基于支持向量机和 XGBoost 的股票评论情感分析：探讨投资者情绪与上证指数的关系及预测

成　　员：　杨穗瑜、吴嘉文、孙玲

学　　院：　信息科学与技术学院（网络空间安全学院）

指导老师：　王连喜

完成时间：　2024 年 12 月 8 日

目录

摘要

随着全球经济和金融市场的迅速发展，投资者情绪作为一种非理性的心理因素，对股票市场表现的影响日益受到重视。情绪分析已逐渐成为金融研究的热点领域，尤其在中国复杂的股票市场中，投资者情绪波动被认为是影响市场波动性的重要因素。本研究以探讨投资者情绪与上证指数的关系为目标，通过数据挖掘和机器学习技术，实现对股评数据的情绪分类和市场预测。

本研究首先爬取了 43 万条来自东方财富股吧的股票评论数据，覆盖了消费、能源、金融、健康、工业和科技六个行业，以及上证指数相关的评论数据。通过使用自然语言处理技术，包括中文分词（基于 jieba）和正则表达式清洗，结合高质量的情感词典（姚加权等 2021 年构建），对股评数据的情绪进行分类和量化。积极评论和消极评论分别通过情感极性值累计计算而得。随后，通过 TF-IDF 特征提取将文本数据向量化，确保了模型训练中输入特征的高质量。

情绪分类任务采用支持向量机（Linear SVC）模型。通过优化正则化参数 C 值，模型实现了 99.49%的最佳分类准确率。情感指数（BI 指数与 BI_Simple 指数）的构建进一步量化了每日情绪波动，用以探讨其与上证指数收盘价的关联性。在情绪波动与市场表现的回归预测中，选择了 XGBoost 模型，结合滞后变量以捕捉历史数据对当前收盘价的潜在影响，显示出对上证指数收盘价的高精度预测能力。

实验结果表明，情感指数与上证指数之间存在显著相关性。情感指标 BI 指数的波动能够反映投资者情绪的剧烈变化，与股市的涨跌趋势高度一致。此外，行业间情绪波动表现出差异性，例如，科技行业的积极情绪与市场创新热点的推动有关，而工业行业则更多受到宏观经济政策的影响。本研究同时发现，当股市表现出现显著变化（如快速上涨或下跌）时，情绪波动也会加剧，进一步验证了情绪波动与市场波动的相互影响。

本研究不仅在情绪分析与市场预测中验证了 Linear SVC 和 XGBoost 模型的有效性，还通过对不同行业股评情绪的细粒度分析，为市场情绪监测和预测工具提供了实践参考。

关键词： 上证指数；支持向量机；XGBoost；情感分析；文本挖掘；股市预测

Abstract

As the global economy and financial markets develop rapidly, investor sentiment, as an irrational psychological factor, is increasingly recognized for its impact on stock market performance. Sentiment analysis has gradually become a hot topic in financial research, especially in China's complex stock market, where investor sentiment fluctuations are considered a significant factor affecting market volatility. This study aims to explore the relationship between investor sentiment and the Shanghai Composite Index (SSE) by employing data mining and machine learning techniques to classify stock review sentiments and predict market trends.

This study collected 430,000 stock reviews from the East Money Stock Bar, covering six industries: consumer goods, energy, finance, healthcare, industrials, and technology, along with comments related to the SSE. Using natural language processing (NLP) techniques, including Chinese word segmentation (based on jieba) and regular expression cleaning, combined with a high-quality sentiment lexicon (constructed by Yao Jiaquan et al., 2021), the study classified and quantified stock review sentiments. Positive and negative reviews were calculated by aggregating sentiment polarity values. Subsequently, the text data were vectorized using TF-IDF feature extraction to ensure high-quality input features for model training.

The sentiment classification task was performed using the Support Vector Machine (Linear SVC) model. By optimizing the regularization parameter CC, the model achieved a best classification accuracy of 99.49%. Sentiment indices (BI Index and BI_Simple Index) were constructed to quantify daily sentiment fluctuations and investigate their correlation with the SSE closing price. For the regression prediction of sentiment fluctuations and market performance, the XGBoost model was selected, incorporating lagged variables to capture the potential influence of historical data on current closing prices, demonstrating high precision in predicting the SSE closing price.

Experimental results show a significant correlation between sentiment indices and the SSE. The fluctuations of the sentiment indicator, BI Index, reflect dramatic changes in investor sentiment, which are highly consistent with stock market trends. Moreover, sentiment fluctuations vary across industries; for example, positive sentiment in the technology sector is driven by innovation hotspots, while the industrial sector is more influenced by macroeconomic policies. The study also found that when the stock market experiences significant changes (e.g., sharp rises or falls), sentiment fluctuations intensify, further verifying the interplay between sentiment and market volatility.

This research not only validates the effectiveness of the Linear SVC and XGBoost models in sentiment analysis and market prediction but also provides practical references for monitoring and forecasting market sentiment through detailed analysis of sentiment in different industry reviews.

Key words： Shanghai Composite Index; Support Vector Machine; XGBoost; Sentiment Analysis; Text Mining; Stock Market Prediction

第一章 绪论

1.1 研究背景

随着全球经济的迅速发展和金融市场的不断扩展，投资者情绪对股票市场表现的影响逐渐受到广泛关注。投资者情绪是指投资者对市场表现的情感和预期，是一种非理性的心理因素，**常常导致股市的波动性增大**。尤其在信息爆炸的时代，新闻报道、社交媒体和市场事件迅速传播，极易引发投资者情绪波动。近几年来，随着数据挖掘、机器学习等技术的发展，情绪分析逐渐成为金融研究的热点领域。研究表明，**投资者情绪对股市的涨跌有显著影响**，其影响力已逐渐持平单纯的市场基本面和技术分析因素。特别是在中国股票市场中，投资者的组成结构复杂，中小投资者的情绪波动更加频繁，而这些情绪波动可能**会放大市场的反应，直接或间接影响上证指数的表现。**如何精确量化投资者情绪与市场表现的关系成为了一个亟待解决的问题。因此，研究投资者情绪与股票市场表现之间的关系，不仅能够更好地理解**市场波动的内在机制**，也为政策制定者和投资者提供了重要的参考价值。

1.2 目的和意义

当前，有关投资者情绪与股票市场表现关系的研究主要集中在**情绪分析和市场预测**两个方向。在情绪分析领域，主流方法包括基于词典的情绪分析和基于机器学习的情绪分类等。词典方法通过构建情绪词汇表，统计文本中的正负面情绪词以生成情绪指标，但这种方法容易受到词典质量和更新周期的限制，难以应对复杂多变的大量文本数据。相比之下，机器学习方法通过对文本数据进行标注和模型训练，能够更灵活地适应不同的文本环境，尤其在社交媒体数据处理中展现出更优越的性能。

在本研究中，我们**结合分类和回归模型，全面分析了投资者情绪与股票市场表现之间的关联性**。首先，**采用支持向量机（SVM）中的 Linear SVC 模型，对投资者情绪数据进行分类分析**。Linear SVC 是一种线性支持向量机分类器，适用于处理线性可分的数据，同时在大规模数据集上表现出较高的计算效率和稳定性。鉴于股票市场波动具有一定的线性特征，该模型能够有效地量化情绪信息，为市场情绪分析提供了坚实的工具支持。

此外，为进一步探讨情绪波动对市场表现的量化影响，我们引入了 **XGBoost 回归模型，以分析情绪数据与上证指数收盘价之间的关联性**。XGBoost 是一种基于梯度提升的集成学习方法，擅长处理非线性关系和高维特征数据，在金融市场复杂性和多样性背景下展现了优异的性能。通过构建滞后变量，该模型能够捕捉投资者情绪的动态变化对市场波动的潜在影响。

研究目标包括以下几个方面：

量化情绪波动对股票市场的实际影响：通过构建滞后特征，分析情绪波动对收盘价的影响；

验证模型性能：利用指标（如均方误差 MSE、均方根误差 RMSE、平均绝

对误差 MAE 和拟合优度 R^2) 评估 XGBoost 模型的预测能力，并通过可视化对比预测值和实际值；

提供高效的情绪分析工具：结合分类和回归分析，建立高效且精确的情绪与市场表现关系模型。

研究结果表明，Linear SVC 模型能够有效捕捉情绪特征的线性关系，而 XGBoost 模型则在处理非线性波动和多维特征方面具有显著优势。通过结合两种模型的方法，本研究不仅拓**展了情绪分析在金融预测中的应用**，还为投资者和政策制定者**提供了识别和管理市场情绪波动的理论支持与实践工具**，同时为后续深度学习在金融市场预测中的应用研究奠定了基础。

1.3 国内外研究现状

在当今金融市场，股票价格预测一直是投资者和研究者关注的焦点。随着信息技术的飞速发展，金融市场数据呈现出多源、异构和海量的特点，这为股票市场预测带来了新的挑战和机遇。在这样的背景下，情感分析作为一种能够从非结构化文本数据中提取情绪信息的技术，逐渐成为金融市场预测的重要工具。

股票市场预测的传统方法，如时间序列分析、回归模型等，往往侧重于量化数据本身，而忽略了投资者情绪这一重要因素。投资者情绪作为一种非理性因素，对股票价格的波动有着不可忽视的影响。因此，结合情感分析技术，研究者们致力于提高股票价格及市场趋势预测的精度，以期为投资者提供更为科学的决策依据。洪巍和李敏[1]在其综述中详细分析了情感词典方法和机器学习方法在文本情感分析中的优缺点，指出了深度学习技术在未来研究中的重要发展方向。

林培光等[2]提出的 SCONV 模型，利用卷积 LSTM 和情感分析相结合的方法，在较小的样本数据集上依然表现出色，通过股吧评论提取股民情绪，显著提升了预测的稳定性和准确性。类似地，许雪晨等[3]提出基于金融文本情感分析的 SA-BERT-LSTM 模型，将 BERT 应用于财经新闻情感分析，并结合股票交易数据，大幅度提高了股指的预测准确率。

其他学者也采用了情感分析技术与不同模型的结合。基于支持向量机 (SVM) 的研究，如 Kumar 等人[4]使用 VSM 情感分类模型来克服传统方法在短文本分析中的不足。李洋和董红斌[5]提出了结合 CNN 与 BiLSTM 的特征融合模型，改善了传统 SVM 的局限性。此外，部慧[6]等采用朴素贝叶斯模型结合股评数据，探讨了情绪对股市表现的影响，表明股评情绪对开盘价、收盘价和交易量的显著预测力。

一些研究通过创新技术进一步提升了情感分析的准确性和实时性。Rizinski 等[7]提出的 XLex 模型结合 Transformer 和 SHAP 工具，扩展情感词典的同时提高了金融情感分析的解释性和实时性，为市场趋势预测提供了参考。Peivandizadeh 等[8]通过融合社交媒体情感和股市数据，提出了结合 Off-policy PPO 算法的 TLSTM 模型，有效解决了情感分析中的不平衡分类问题，并在多项指标上实现最佳预测效果。

通过整合多种方法和技术，这些研究表明，将情感分析与传统预测模型相结合，能够显著提升股票市场预测的精度与稳定性。然而，模型复杂度高、对大规模数据的依赖以及情感特征提取的准确性仍是需要进一步解决的关键问题。未来研究可以通过优化模型架构，增强情感分析的实时性和准确性，从而为投资者和市场决策提供更有效的支持。

1.4 数据挖掘的主要方法

1.4.1 分类分析

分类分析是处理数据分类问题的一种数据分析方法。与数值分析不同，数值分析侧重于处理连续型数据的数值关系，而分类分析主要聚焦于对具有不同特征的数据进行分类。它依据数据的属性、特征等将数据划分为不同的类别。在数据挖掘和机器学习领域，研究数据分类可以运用分类分析以及聚类分析（clustering analysis）。

1.4.2 回归分析

回归分析是处理多变量间相关关系的一种数学方法。相关关系不同于函数关系，后者反映变量间的严格依存性，而前者则表现出一定程度的波动性或随机性，对自变量的每一取值，因变量可以有多个数值与之相对应。在统计上研究相关关系可以运用回归分析和相关分析（correlation analysis）。

1.5 数据挖掘的过程

1.5.1 确定研究目的及对象

在股评的情感指数和上证指数的相关性的研究中，涉及的各行业的股评以及上证指数的数据，应从当日股评情感和当日的上证指数的关系的的出发，目的是挖掘出股评情感的走势对上证指数的影响，以及影响较大的行业，并以此为依据，选择适合目标对象的数据挖掘方法，例如**基于支持向量机和 XGBoost**。

1.5.2 数据准备

建立与之相关的数据流，比如 CSV 文件或者数据库，并搜集可靠的数据，载着我们我们爬取了东方财富吧的**股评**，以及**上证指数**作为数据源。由于原始数据涉及的信息不一定完全满足数据挖掘的需要，因此需要对数据进行选择和处理，使数据转化为构建模型所需的格式，并对数据做必要的文字描述。

1.5.3 数据预处理

主要包括数据清洗、特征构造和特征选择等几个过程。数据清洗的目的是补全数据、处理缺失值、除去噪声以及改正不协调的数据。在这里我们对不全的数据进行删除和补全。数据转换主要包括构造新的衍生特征和对连续型数据进行规

范化。通过特征选择可以减少样本的维度，大大减少计算量，降低时间和空间复杂度，简化模型。

1.5.4 建模过程

根据数据挖掘的目的以及数据的特点，选择相应的数据挖掘方法和算法，建立模型，并对经过预处理的数据进行挖掘。我们选择基于支持向量机和 XGBoost 来对**评论的情感指数和上证指数的相关性**进行建模。

1.5.5 结果分析

根据数据的现实意义，结合数据挖掘的目的，分析和解释数据挖掘的结果，并做出相应的效果评估和具体的相关性描述。

1.5.6 模型评估与调整优化

由于选取的数据可能存在一定的偶然性和必然性，不能保证挖掘出的结果就是正确使用的，因此需要对挖掘出的模型进行评估和检验。在评估和检验的分析结果基础上对模型进行调整和优化，以保证所挖掘的结果更有效、更使用，更能准确的反映出股评情感指数和上证指数之间的关系。

第二章 应用数据挖掘方法研究投资者情绪与股市表现关系并预测

在本研究中，如图 1 所示，我们结合 Linear SVC 分类模型和 XGBoost 回归模型，分析投资者情绪与股票市场表现的关联性。通过数据预处理、TF-IDF 特征提取，预测情感极性并计算情感指标，最终揭示情感变化与上证指数波动之间的关系。

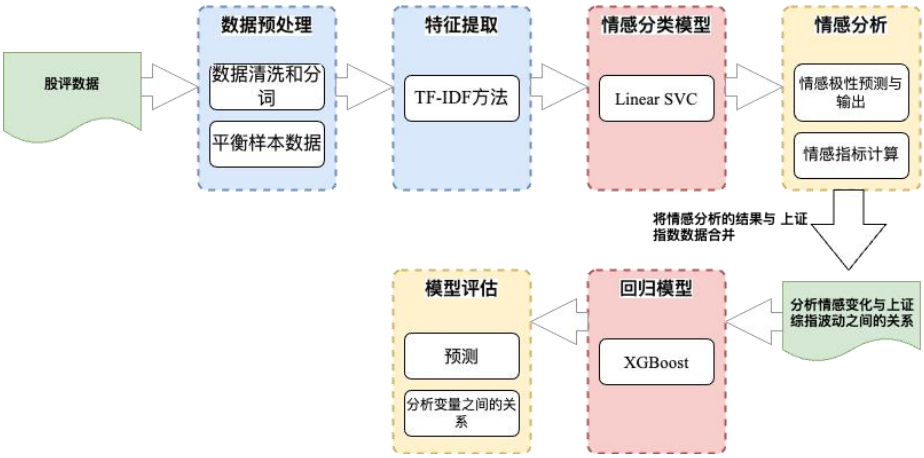


图 1 数据挖掘流程图

2.1 确定挖掘的数据对象

本次共从东方财富股吧爬取了 43 万条的**七月到九月**的**股评数据**。包括了**消费、能源、金融、健康、工业、科技**六个行业以及**上证指数**的评论，其中，包括**评论的时间，评论的内容**，如图 2 所示，我们爬取了健康、能源、金融行业部分股评数据。

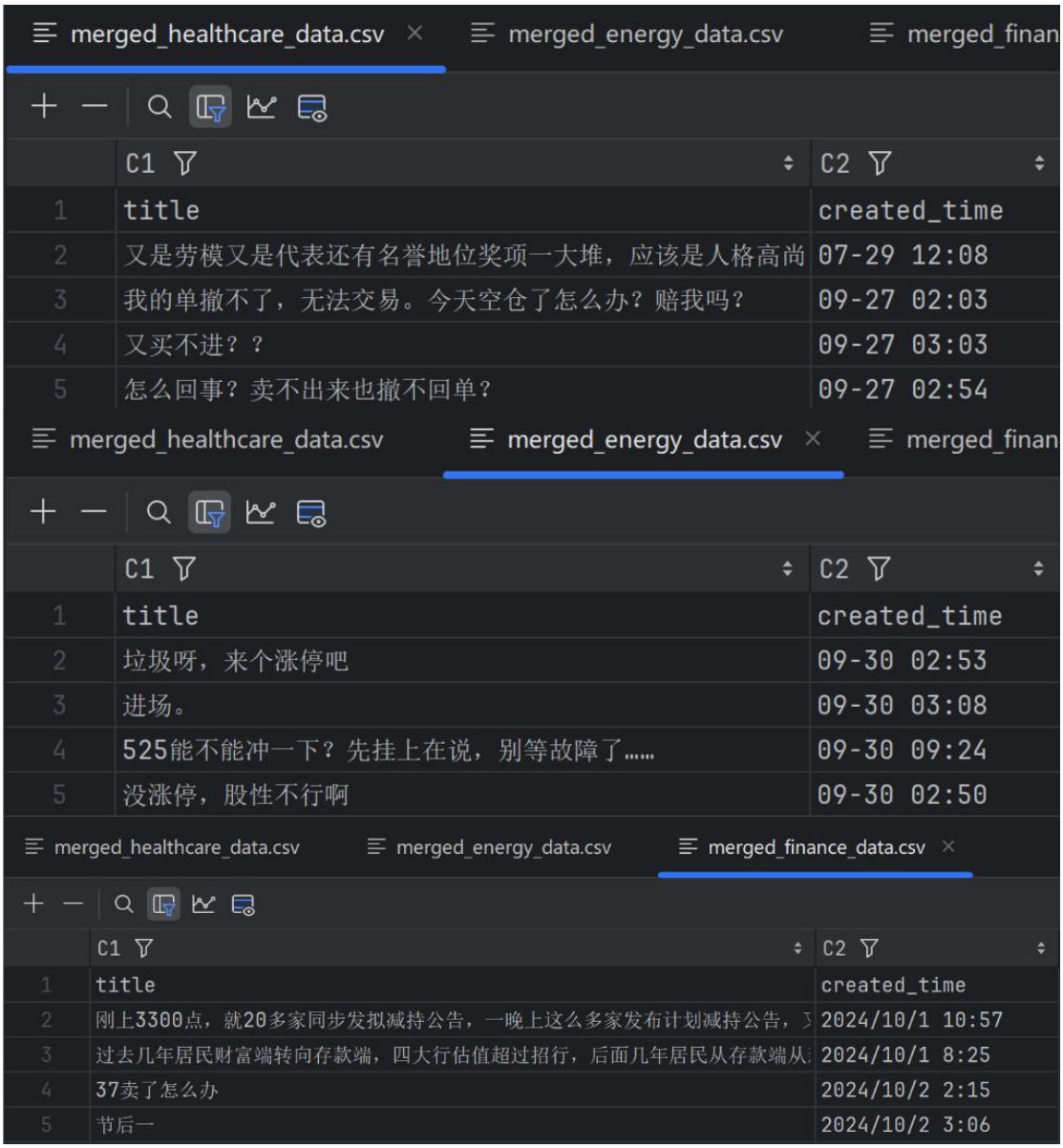


图 2 健康、能源、金融行业部分股评数据

同时，我们还爬取了七月到九月的上证指数的日收市情况数据，其中包括了**日期，开市价、最高价、最低价、闭市价**以及**交易量**的六个特征，如图 3 所示，我们爬取了部分上证指数数据。

date	open	high	low	close	volume
2024-10-31	3,279.82	3,267.98	3,295.74	3,252.39	78.46B
2024-10-30	3,266.24	3,273.64	3,291.68	3,244.81	66.90B
2024-10-29	3,286.41	3,328.09	3,340.46	3,284.21	72.53B
2024-10-28	3,322.20	3,300.46	3,322.20	3,279.72	67.24B
2024-10-25	3,299.70	3,280.76	3,319.36	3,276.13	59.57B
2024-10-24	3,280.26	3,287.82	3,292.94	3,266.88	51.98B
2024-10-23	3,302.80	3,285.25	3,331.08	3,277.07	65.04B
2024-10-22	3,285.87	3,263.82	3,294.96	3,255.14	57.49B
2024-10-21	3,268.11	3,276.06	3,300.66	3,239.10	66.82B
2024-10-18	3,261.56	3,165.97	3,313.98	3,152.82	69.30B

图 3 部分上证指数数据

在所采集的数据中，包含了**数值型和非数值型两类数据**，其中，股评的日期和评论内容以及上证指数的日期为非数值型数据，**开市价、最高价、最低价、闭市价**都为数值型数据

为探讨股评的情感指数和上证指数之间的关系，我们选取**闭市价**作为**当日最终的上证指数**，再通过对评论的内容进行情感划分以及情感指数计算进行研究。

2.2 数据预处理

在进行评论内容的情感量化分析之前，我们首先采用**自然语言处理**（NLP）中的**分词**技术对评论文本进行粒度化处理，这一步骤是**将连续的文本流分解成可管理的词汇单元**，在这里我们使用 jieba **分词技术**，这是一个专门为中文设计的分词库，并且使用了**正则表达式来去除标点符号** ，如图 4 所示，部分分词后的股评数据。

C1	C2
title	created_time
卖 都 卖不出去 太 离谱 了	2024/09/27 02:15
垃圾 产品	2024/09/27 09:45
上涨 空间 完全 打开 了	2024/09/27 06:06
慢慢来 不要 飞 的 太快 大笑 注重 长线 价值 爆涨 爆跌 不利于 企业形	2024/09/27 09:17
买 进去 100 股 即 没 看见 卖出 也 没 看见 钱 回来 反正 是 股票 没	2024/09/27 03:04
大佬 们 这 今天 大盘 涨 的 86 个点 这 也 对不上 有 很多 没有 公布	2024/09/27 05:27

图 4 部分分词后的股评数据

接下来，我们引入**情感词典**的概念。情感词典是情感分析任务的核心资源，它**包含了词汇与其情感值之间的映射关系，用于评估文本中词汇的情感倾向**。构建一个高质量的情感词典对于提升情感分析的准确性和效果具有决定性作用。在本研究中，我们采用了**姚加权等 (2021) 构建的股吧社媒情感词典**。姚加权等 (2021) 构建的股吧社媒情感词典的部分数据如图 5 所示。

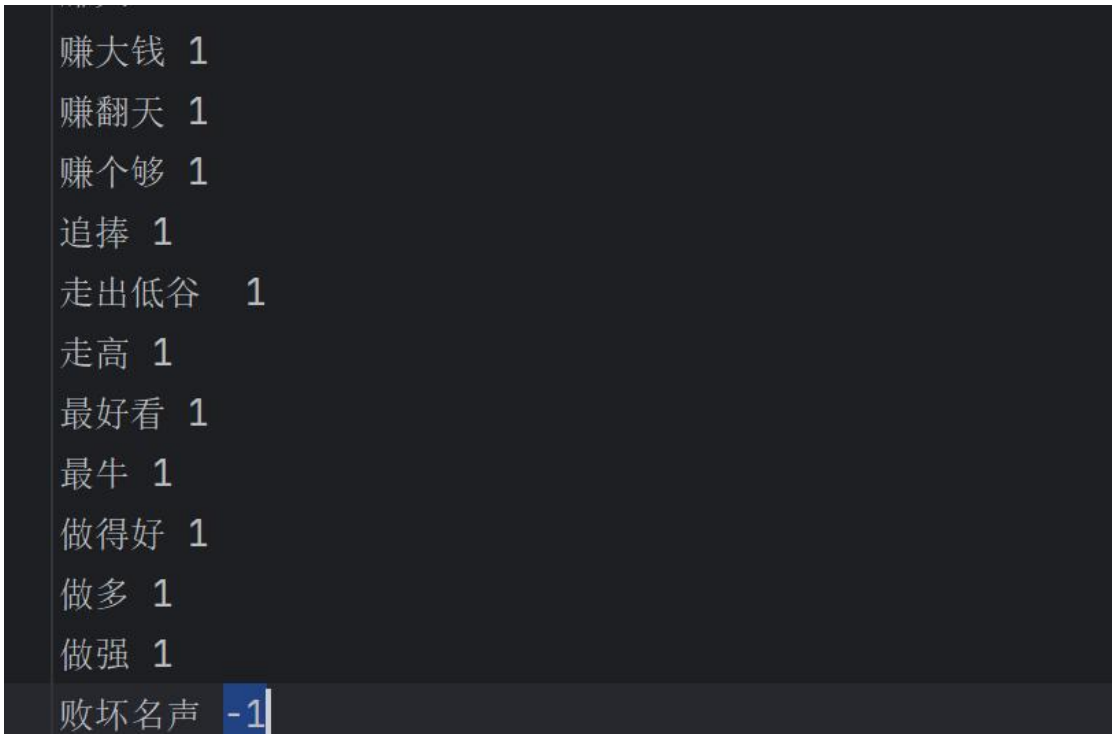
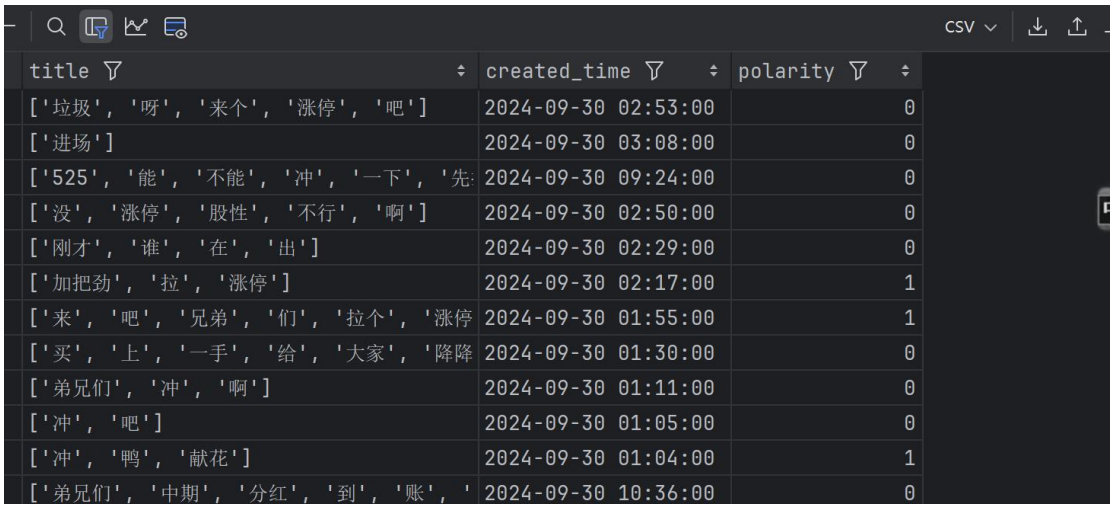


图 5 部分姚加权等（2021）的股吧社媒情感词典的数据展示

分词完成后，我们**利用情感词典对句子中的词汇进行情感极性评分**，将**积极情感词汇的极性强度赋予正值 1，消极情感词汇的极性强度赋予负值 1**。随后，我们将句子中所有词汇的情感极性值进行累加，以此作为情感倾向的量化指标。如果该**综合得分大于等于 0，则将该评论定性为积极评论；如果得分小于 0，则定性为消极评论**，股评数据在情感定性后的部分数据如图 6 所示。



CSV

title	created_time	polarity
['垃圾', '呀', '来个', '涨停', '吧']	2024-09-30 02:53:00	0
['进场']	2024-09-30 03:08:00	0
['525', '能', '不能', '冲', '一下', '先	2024-09-30 09:24:00	0
['没', '涨停', '股性', '不行', '啊']	2024-09-30 02:50:00	0
['刚才', '谁', '在', '出']	2024-09-30 02:29:00	0
['加把劲', '拉', '涨停']	2024-09-30 02:17:00	1
['来', '吧', '兄弟', '们', '拉个', '涨停	2024-09-30 01:55:00	1
['买', '上', '一手', '给', '大家', '降降	2024-09-30 01:30:00	0
['弟兄们', '冲', '啊']	2024-09-30 01:11:00	0
['冲', '吧']	2024-09-30 01:05:00	0
['冲', '鸭', '献花']	2024-09-30 01:04:00	1
['弟兄们', '中期', '分红', '到', '账', '	2024-09-30 10:36:00	0

图 6 情感定性后的部分股评数据展示

在获得了句子的情感标签之后，我们进一步**将文本数据转换为数值向量**，以便用于机器学习模型的训练。在此过程中，我们采用了 **TF-IDF（Term Frequency-Inverse Document Frequency）向量化方法**，通过 TfidfVectorizer 将文本数据转换为 **TF-IDF 向量**，这一步骤能够突出重要词汇，降低常见词汇的权重，从而为**线性支持向量机（LinearSVC）**模型的训练提供高质量的特征表示。

2.3 描述性分析

2.3.1 积极消极情感词汇的频率分布分析

通过分析不同行业（**工业**、**消费**、**能源**、**金融**、**医疗保健**、**科技**、**上证指数**相关）2024 年 7 月至 9 月股评数据中的情感词汇，统计**积极**和**消极**情感词汇的频次，绘制柱状图来直观展示每个行业股评中前 10 个积极和消极情感词汇的频率分布情况，同时从中可以获得一些关键信息。

以上证指数股评为代表（见图 7 和图 8），其中消极和积极的前十个词汇分别围绕“**涨**”和“**跌**”，反映出股民最关心的即股票的涨跌。通过两图的对比，“跌”为代表的消极词汇多于以“**涨**”为代表的积极词汇，暗示这段时间内市场情绪较为**悲观**。

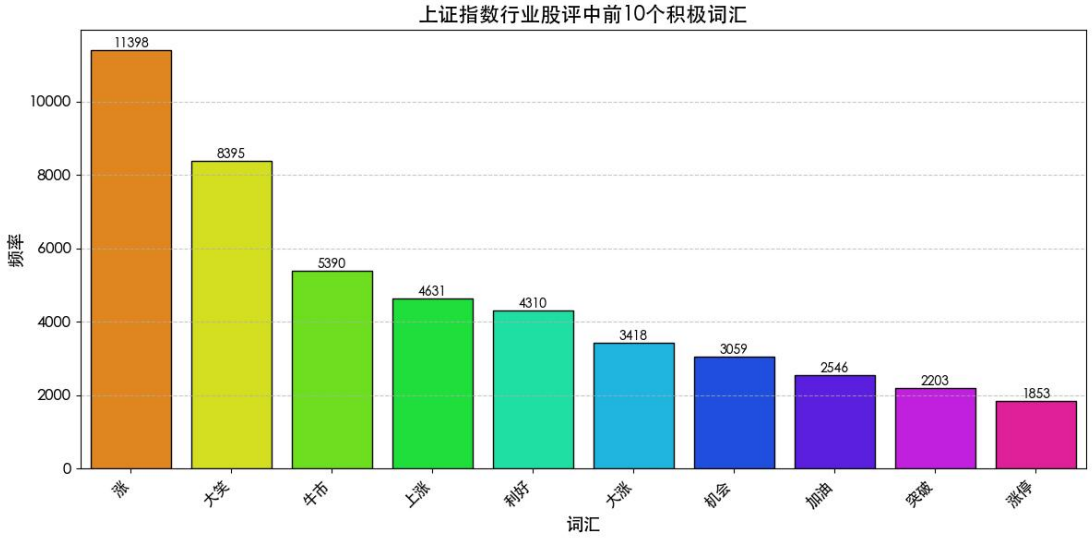


图 7 上证指数行业股评中前 10 个积极词汇

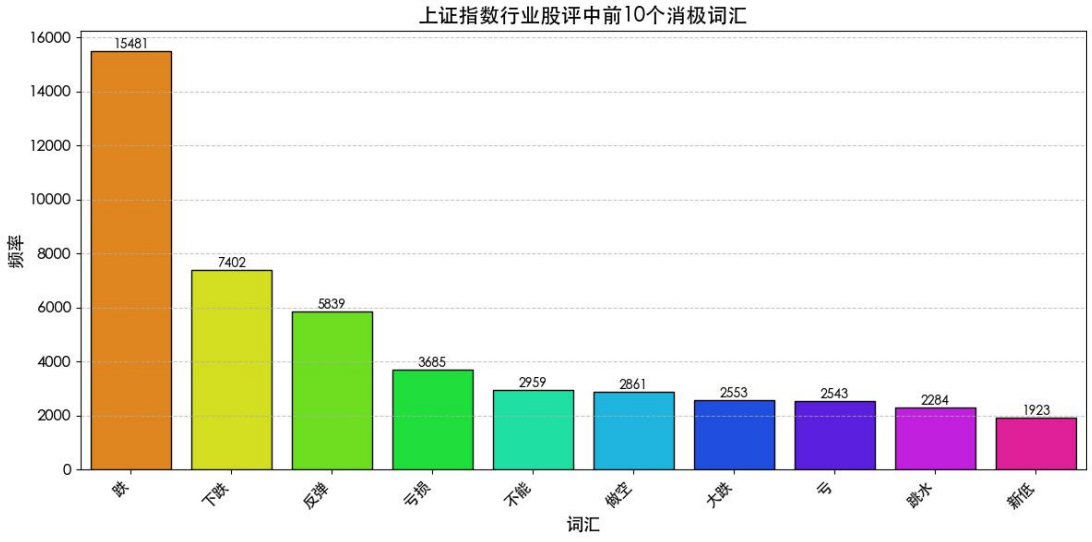


图 8 上证指数行业股评中前 10 个消极词汇

其他各行业类似上证指数股评，通过比较积极和消极情感词汇的柱状图可以揭示该行业股评中的整体情感倾向。如果在行业中，**积极情感**词汇的频率总和高于**消极情感**词汇的频率总和，这可能表示行业的股评整体上更倾向于**积极评价**；反之，如果消极情感词汇频率更高，则表明股评对行业更多持有消极态度。

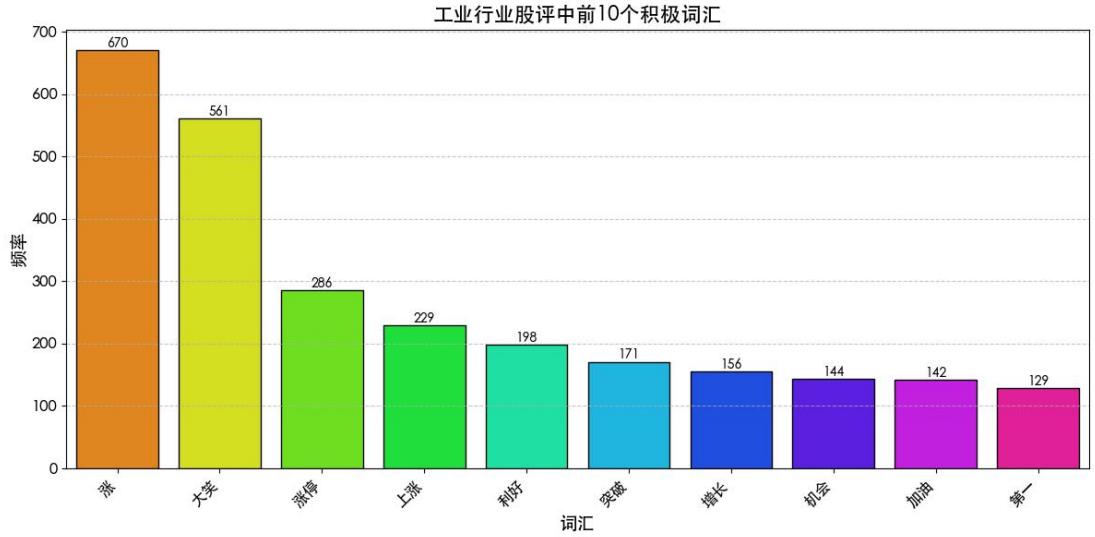


图 9 工业行业股评中前 10 个积极词汇

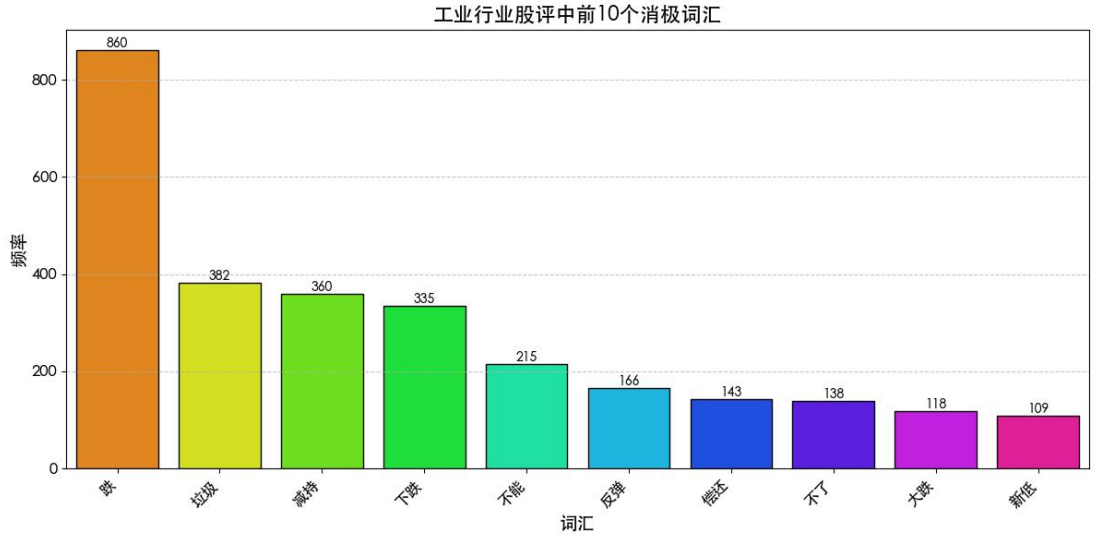


图 10 工业行业股评中前 10 个消极词汇

如**图 9** 和**图 10**，在工业行业中消极情感词汇频率更高，则表明投资者对工业行业更多持有**消极**态度。

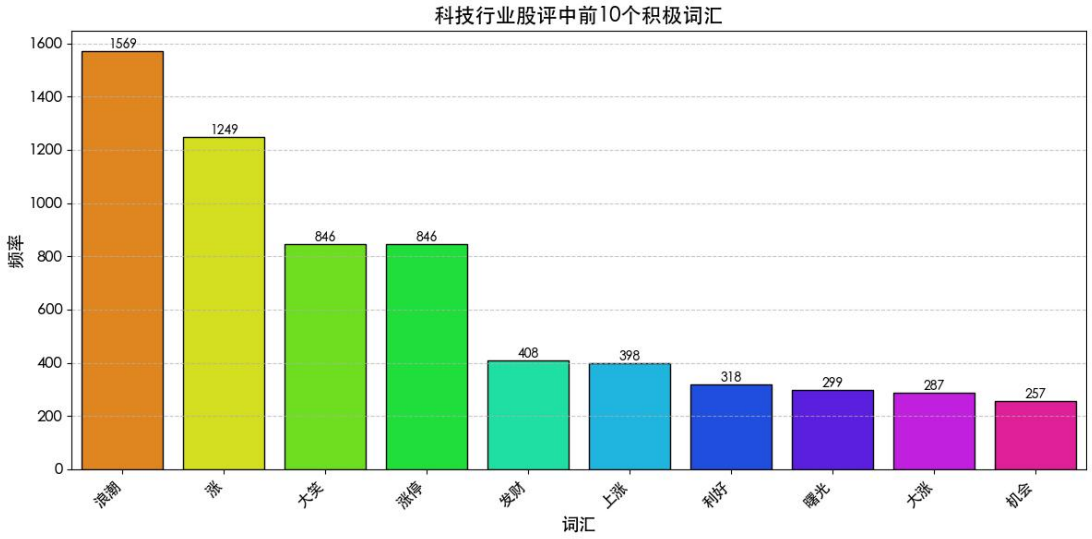


图 11 科技行业股评中前 10 个积极词汇

通过查看**不同行业**的积极和消极情感词汇柱状图，可以比较不同行业股评中的**情感倾向**差异。例如，对比**工业**和**科技**行业的积极情感词汇频率分布（见**图 9** 和**图 11**），工业行业的积极情感词汇频率普遍**较低**，而科技行业相对**较高**，这可能暗示在股评中科技行业的正面评价更多或者积极情感表达更强烈。

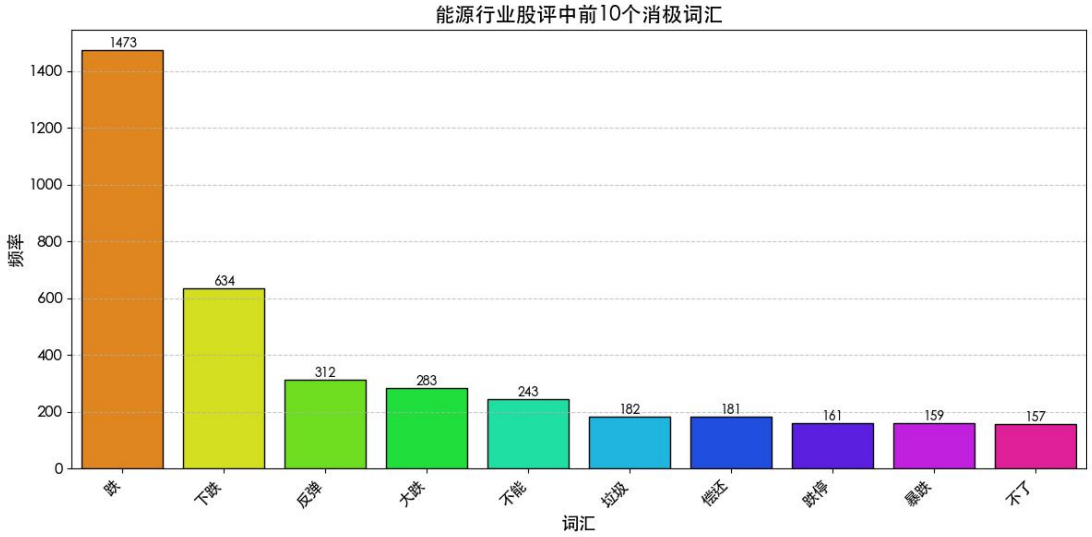


图 12 能源行业股评中前 10 个消极词汇

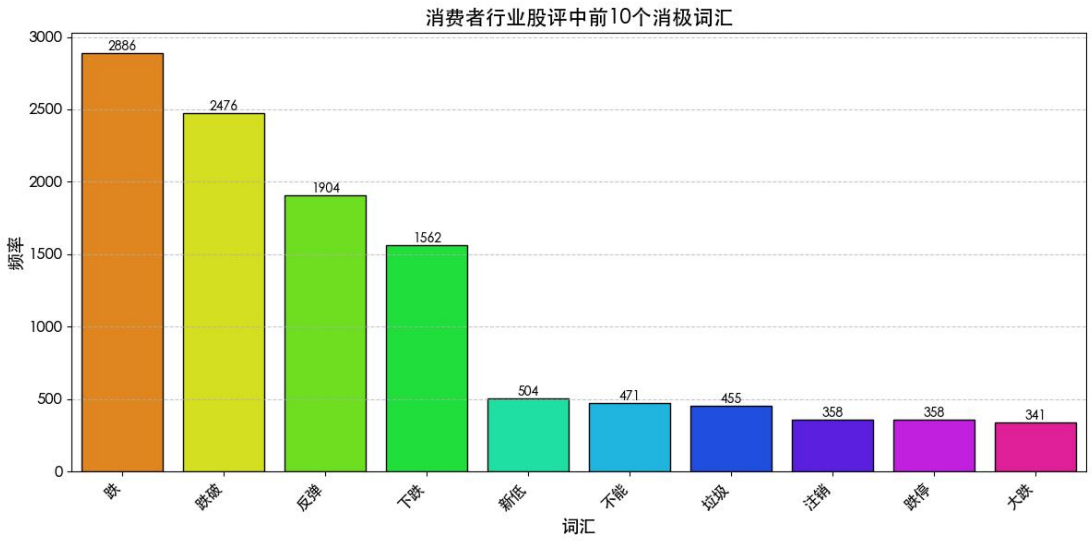


图 13 消费者行业股评中前 10 个消极词汇

对于**消极**情感词汇，发现**消费者行业**的消极情感词汇频率波动较大，而**能源行业**较为稳定（如**图 12** 和**图 13**），这可能反映出消费者行业的股评情绪更不稳定，更容易出现极端的消极评价，而能源行业的消极评价相对较为温和且一致。

2.3.2 各行业词云图分布分析

词云图中，词汇的**大小**反映了其在原始文本中的**出现频率**。出现频率越高的词汇，在词云图中的字体越大，越居于**中心**位置；而出现频率较低的词汇则字体较小，位置相对**边缘**。例如，如果在词云图中“上涨”这个词的**字体**较大且**位置**较为中心，说明在股评数据中“上涨”这个词出现的**频率**相对较高，可能是股评中经常提到的一个**概念**或者是描述上证指数走势的关键**内容**。

通过观察词云图中的词汇，可以快速了解到**股评数据**中涉及的主要概念和话题。比如可能会看到与股市相关的词汇如“股票”“指数”“行情”等，以及与经济形势或企业表现相关的词汇，这些词汇共同构成了对各行业股评内容的一个概括性的**视觉呈现**。

图 14 上证指数股评词云图

整个词云图是对行业股评内容的主题概括。它以直观的视觉形式展示了股评中最常出现的词汇，从而反映出股评的主要关注点。

例如**图** 14，词云图中出现较多“**市场**”“**银行**”“**散户**”“**行情**”以及上证指数的代码“sh000001”等词汇。这可能意味着市场整体的关注度高，以及**银行**板块的业绩表现、资产质量、利率政策对银行的影响等都会对上证指数产生重大影响，同时**散户**对上证指数的关注度，他们可能是市场中的活跃参与者，其**买卖行为**、**投资情绪**等都会对上证指数产生一定的影响。例如，当散户普遍恐慌**抛售**股票时，可能会导致上证指数**下跌**；而当散户**积极入市**时，可能会推动指数**上升**。

图 15 消费行业股评词云图

而在**图** 15 中，消费行业股评词云图中，“茅台”、“贵州”、“伊利”、“白酒”、“触底”、“反弹”等词汇出现较多，这可能意味着以下几个方面：

1. 行业关注度

这些词汇的频繁出现表明市场对**白酒**行业的关注度较高。白酒作为中国传统

的饮品，具有广泛的消费群体和深厚的文化底蕴，因此在股票市场中也备受关注。

2. 贵州茅台的影响力

“茅台”和“贵州”这两个词汇的出现，可能是因为贵州茅台作为中国白酒行业的龙头企业，其业绩表现、市场份额和品牌影响力都非常大。贵州茅台的股价走势往往能够反映整个白酒行业的**发展趋势**。

3. 伊利的关联性

“伊利”这个词汇的出现可能是因为伊利股份是中国乳制品行业的领军企业，与白酒行业存在一定的关联性。例如，白酒和乳制品都是消费品，它们的消费群体和市场需求可能存在一定的重叠。

4. 触底反弹的预期

“触底”和“反弹”这两个词汇的出现，可能是因为市场对白酒行业的股价走势存在一定的预期。当行业经历了一段时间的**调整**或者**下跌**后，投资者可能预期行业将迎来反弹，因此在股评中会频繁出现这两个词汇。

5. 政策影响

白酒行业的发展受到政策的影响较大，例如税收政策、行业准入政策等。因此，股评中出现这些词汇也可能是因为市场对政策变化的预期或者反应。

综上所述，这些词汇在词云图中的频繁出现，可能是因为市场对**白酒行业**的关注度较高，并且对行业的发展趋势存在一定的预期。投资者和分析师在评估白酒行业的股票时，可能会重点关注这些方面的信息。

图 16 能源行业股评词云图

词云图还可以反映出市场的**情绪倾向**。如果“**跌**”“**低**”等词汇较大且较明显，可能暗示股评整体对行业股票的走势持较为**谨慎**或**悲观**的态度（如图 16）；

图 17 工业行业股评词云图

相反，如果“**上涨**”“**机会**”“**分红**”等词汇突出（如图 17），可能表示股评对市场较为**乐观。**

最后关于词云图的**局限性**，**词云图**虽然能够快速展示文本数据中的主要词汇，但它不能提供词汇之间的**关系信息**，也不能**精确**地显示词汇的频率数值。例如，我们只能通过**字体大小**大致判断词汇的**相对频率**，但无法确切知道某个词汇到底出现了多少次。而且，词云图对文本数据的展示是一种简化的方式，可能会**忽略**一些**低频**但重要的词汇或概念。

2.4 建立模型

2.4.1 情感分类模型训练与优化

在完成特征提取后，股评文本数据通过 TF-IDF 编码被转化为特征矩阵 X，其中每一行代表一个股评，每一列代表一个词汇的权重。对应的情感标签被转化为目标变量 y，其中 1 表示正面情感，0 表示负面情感。

LinearSVC 是 Scikit-learn 中的一个分类模型，用于解决线性可分的分类问题。它基于支持向量机（Support Vector Machine, SVM）算法，但采用线性核函数。本研究调用 **sklearn.svm.LinearSVC 类**建立一个线性支持向量机（SVM）模型，通过训练数据 X 和情感标签 y，模型学习每个特征（即词汇）在情感分类中的权重和重要性。这个过程中，SVM 会根据每个词汇的 TF-IDF 权重来评估它对正面或负面情感的贡献，进而构建一个能够区分正面与负面情感的决策边界。

为了优化分类器的性能，我们通过 网格搜索（GridSearchCV） 调整模型的 C 参数，即正则化参数，它控制着模型的复杂度和拟合能力。通过在不同的 C 值下训练模型，网格搜索会评估每个参数组合在交叉验证中的表现，最终选择出能够在验证集上达到最佳准确率的参数。这样，我们能够获得最适合当前任务的模型，从而提高情感分类的准确性。在最佳准确率下使用的 C 值如表 1 所示

表 1 参数 C 最优值与最佳准确率

C 值	准确率
10	0.9949

在模型训练完成后，利用训练好的分类器对新的股评数据进行情感预测。首先，通过相同的 TF-IDF 特征提取方法将新的股评文本数据转化为特征矩阵。然后，使用训练好的 SVM 分类器对新的数据进行预测，并将预测结果量化为二元情感标签，其中 1 表示正面情感，0 表示负面情感。通过这种方式，模型能够自动识别并分类新的股评的情感倾向。

2.4.2 情感指数计算

在情感分类结果的基础上，我们通过日期分组进一步构建情感指数来分析情感波动对股票市场表现的影响：

BI 指数（Log 比例指数）：对于每一天的数据，统计正面评论数和负面评论数，代入公式计算情感指数。1 的加法项可避免当某一类评论数为零时分母取值错误。

$$BI=log(\frac{1+正面情感股评数量}{1+负面情感股评数量})$$

BI_Simple 指数（简单差值指数）：使用每日正负评论数的差值与总数的比值，直接反映情感的倾向性。

$$BI\_simple=\frac{正面情感股评数量-负面情感股评数量}{正面情感股评数量+负面情感股评数量}$$

最后，将 7-9 月期间所有行业的每日情感指数（BI 和 BI_Simple）与上证指数的相关数据进行合并，并对合并后的数据进行可视化分析。以消费行业为例，如图 18 所示。

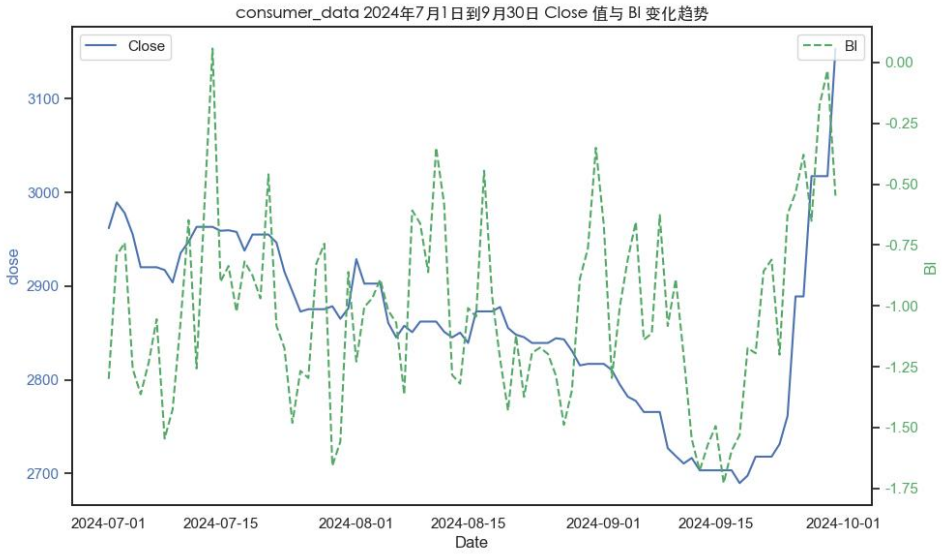


图 18 消费行业情感指数与上证指数的趋势对比

2.4.3 回归模型构建

在回归模型的构建中，通过对原始数据进行系统的预处理和特征工程，设计

了一个基于 XGBoost 的回归预测框架，用于预测上证指数的收盘价。针对时间序列数据的特性，特意引入了滞后特征，这一做法对于时间序列预测至关重要，因为当前的股市价格通常受历史价格和交易数据的影响。通过对包括开盘价、最高价、最低价、交易量及情感指标（BI_Simple）等关键变量构建一到两天的滞后值，我们能够有效捕捉历史数据对当前收盘价的潜在影响。滞后特征的引入不仅丰富了特征空间，使模型能够利用历史信息来预测未来趋势，还增强了模型对时间序列中动态关联和时序性变化的捕捉能力，进一步提升了模型对股市波动规律的适应性和预测准确性。

通过清理缺失值、标准化特征值，以及划分训练集和测试集等步骤，确保了数据的质量与模型的泛化能力。在模型训练环节，选用了强大的 XGBoost 回归器，结合其对非线性关系的建模能力，有效提高了预测的精度和稳定性。

2.4.4 回归模型优化

通过网格搜索（GridSearchCV）对超参数进行优化，搜索范围涵盖了树的数量（n_estimators）、最大深度（max_depth）、学习率（learning_rate）、子采样比例（subsample）以及特征采样比例（colsample_bytree）等关键参数。在此过程中，采用三折交叉验证（CV）和均方误差（MSE）作为优化目标，以确保模型的鲁棒性和稳定性。经过优化，最终得到了最佳参数组合，如表 2 所示。

表 2　XGBoost 模型训练得到的最佳参数组合

每棵树训练时随机选择的特征的比例	**学习率**	**树的最大深度**	**树的数量**	每棵树训练时随机选择的样本比例
0.7	0.1	10	300	0.7

最终，使用最佳参数的模型对测试集进行了预测。在此参数配置下，模型的平均绝对误差（MAE）为 4.67，决定系数（R^2）达到了 0.9937。结果显示，该模型能够较好地预测上证指数的收盘价，具有较高的精度和解释能力。同时，通过可视化对比真实值与预测值的分布，如图 19 所示，进一步验证了模型的优良性能，为金融市场数据的分析与预测提供了有效支持。

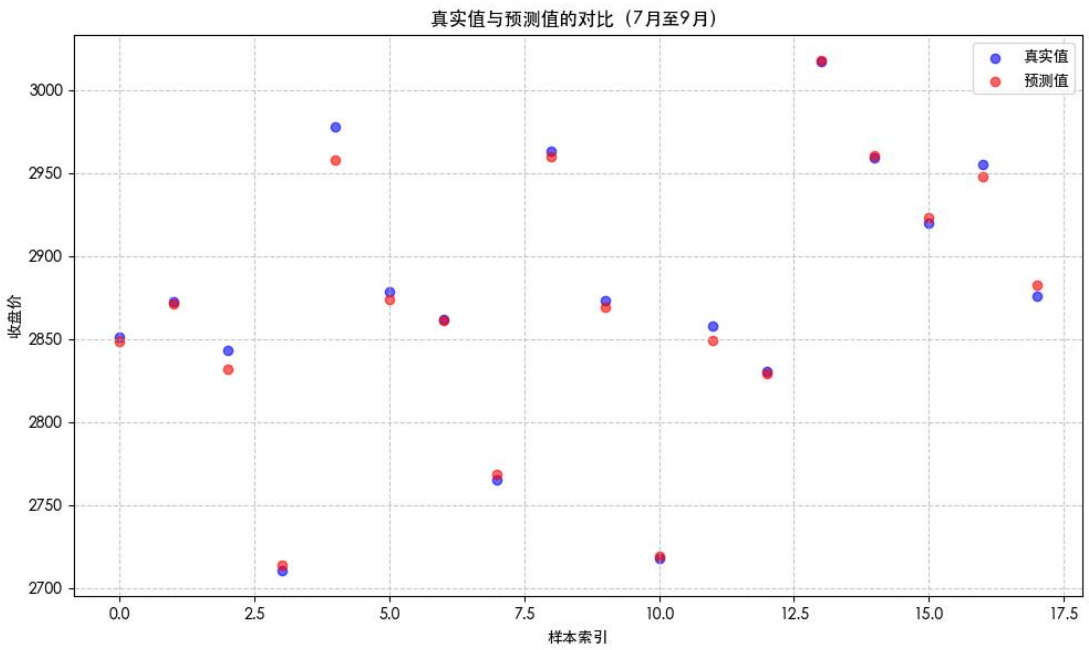


图 19 真实值与预测值的分布

第三章 实验结果与分析

3.1 结论与分析

3.1.1 投资者情绪与上证指数的关系表现分析

通过情感分类模型（SVM）对股评数据进行情感分析，并计算两种情感指数：BI 指数和 BI_Simple 指数。这些情感指数可以用来衡量每日的股评情感波动，并进一步分析其与上证指数（如图 20）的相关性。

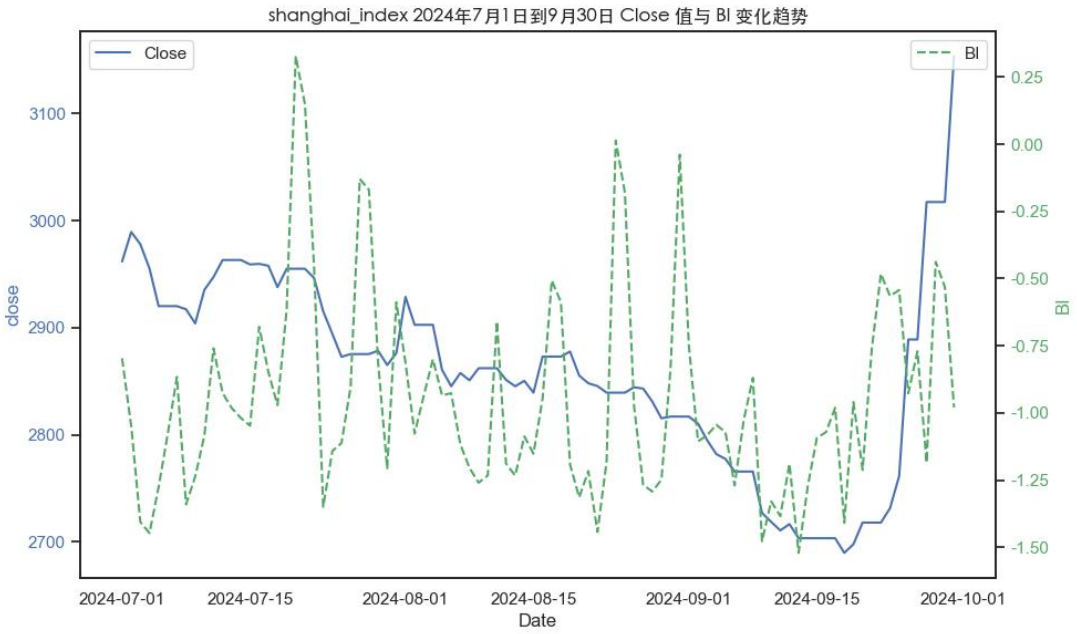


图 20 上证指数 Close 和 BI 变化趋势

1. 情感指数与上证指数波动的关系：

BI 指数的波动能够反映股评的情感波动。当情感指数急剧波动时，可以观察到**上证指数 Close 值**也有明显的变化。例如，在**九月上旬，**情感指数不断下降，上证指数也出现了相应的**下跌**趋势，而在后面的**九月下旬**到**十月初**，情感指数出现上浮波动，同时也观察到上证指数在飞速的**上涨**。

2. 情感波动与股市涨跌的相关性：

正面情感的增加：当情感指数显示正面情感的显著增加时（如 BI 值的上升），可以推测投资者的信心增强，股市可能会迎来短期上涨。

负面情感的增加：当负面情感占主导（BI 值下降），股市可能会出现下跌或波动加剧的现象。

从图中可以看到，当股市有较大波动时（如上证指数的明显上涨或下跌），情感指数通常会有所响应。这表明股评情感可能与股市的短期表现密切相关。

3. 情感指数波动的周期性：

从图中我们可以发现情感指数的波动与某些特定的时间周期（如某些月份或节假日后）相关，说明股市的情感波动可能与周期性因素（如季节性效应、政策发布、财报季等）有一定的联系。

根据 2024 年 7 月到 10 月的上证指数实际表现，我们可以具体分析这些情感波动与实际股市表现的关系。图表中如负面情感的增加，出现在股市下跌期间，可能是因为：

市场对不利政策的反应：政策如房地产调控或货币收紧导致市场不确定性增加，股评的负面情绪加剧，导致情感指数下降，并与股市下跌相关联。

财报季的负面影响：如果大部分上市公司财报不如预期，股民的情绪可能会变得消极，情感指数可能显示出负面波动，这时上证指数可能出现较大幅度的下跌。

而在 2024 年 9 月到 10 月期间，特定的**宏观经济事件**（如**降准降息的货币政策**、下调二套房贷款最低首付比例等**房地产政策**），这些事件通常直接影响了市场的情感波动。这些事件对经济前景的预期较为乐观，股市的情感表现为积极，情感指数的上升可能与股市的上涨形成正向关联。

相反，如果某些不确定性事件（如**国际贸易摩擦、政策收紧、外围市场影响**等）引发市场恐慌，情感指数可能表现为负面波动，股市可能出现调整或回落。

4. 情感指数与股市回调：

在一些股市下跌的期间，可能伴随着情感指数的负向波动。这表明负面情绪的加剧可能会导致市场的回调或进一步下跌。相反，当市场情绪恢复到正面时，股市可能会随之恢复上涨。

5. 行业间的差异：

不同的行业可能会对情感波动的反应不同。例如，消费行业和技术行业的股评情感可能受到不同的市场事件或政策变化影响，因此情感指数与股市表现的相关性存在差异。

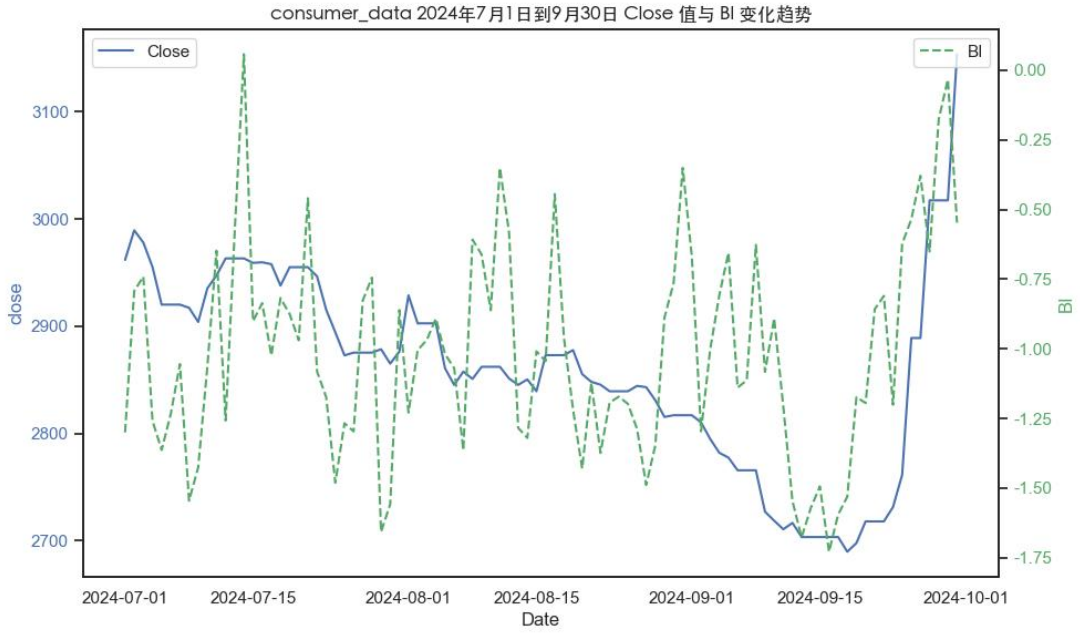


图 21 消费者行业 Close 和 BI 变化趋势

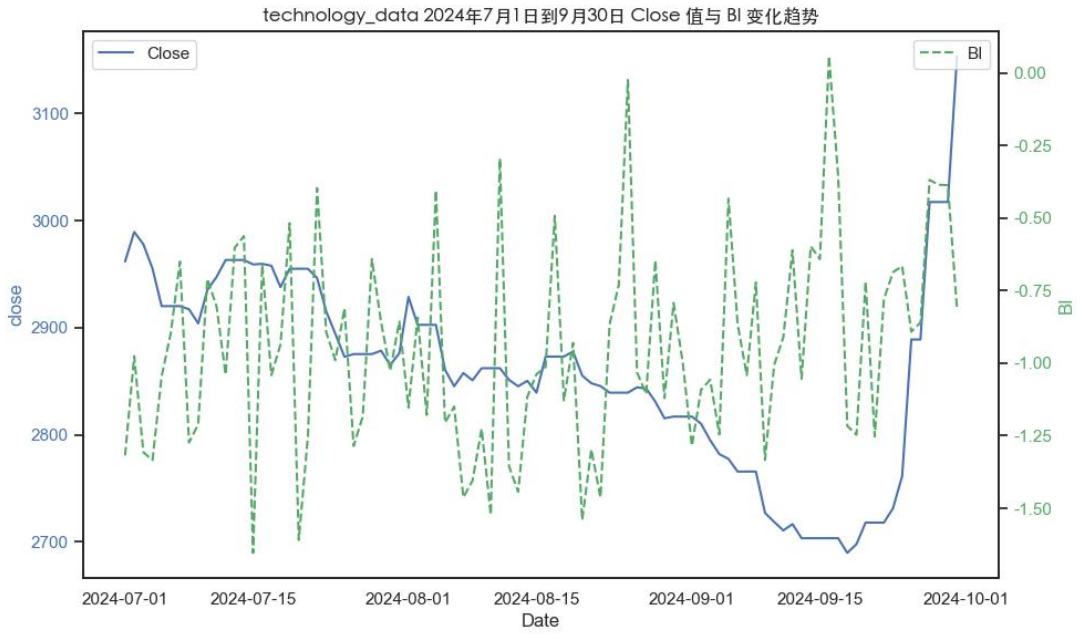


图 22 科技行业 Close 和 BI 变化趋势

如图 21 和图 22，九月初旬消费者行业股民情绪的持续低迷到九月末旬的上涨与股市表现基本成正相关，而对于科技行业，在股市表现持续低迷的情况下，股民情绪却持续上涨，这是由于新兴技术如 AI、半导体等领域的突破带动了相关股票上涨，科技监管政策、创新激励措施等将对科技行业股市产生的影响 。

3.1.2 投资者情绪与上证指数的关系表现回归预测分析

结合股市的实际情况，对实验结果**图 19**（预测值与真实值的对比图）进行以下几个方面的分析：

1. 预测值与真实值的分布关系：

拟合效果：观察回归图中预测值与真实值的分布。可以发现大多数真实值与预测值的点高度吻合，说明模型能够很好地捕捉收盘价的变化规律，预测结果与实际数据之间具有较高的吻合度。

偏差分析：注意到有些数据点偏离斜线较远，这些点可能代表模型预测效果较差的情况。九月末房地产政策、资本市场政策以及宏观经济政策的发布引起股市剧烈波动，模型可能难以捕捉这种复杂的非线性变化。

2. 滞后特征的有效性：

由于模型引入了开盘价、最高价、最低价、交易量以及情感指标 (BI_Simple) 等的滞后特征，回归图可以用来验证这些特征对模型预测能力的贡献。而回归图显示预测值与真实值之间的误差很小，说明滞后特征在捕捉历史对当前的影响上是有效的。

具体而言，情感指标（BI_Simple）的滞后特征可能反映了投资者情绪对股市短期趋势的延续性影响。如果市场情绪（如大量正面评论）与股市上涨具有显

著相关性，这一特征的引入会进一步提高预测精度。

3. 模型的稳定性与泛化能力：

决定系数 $R2=0.9937R^2 = 0.9937$ 表明模型对训练数据的解释能力非常强。结合回归图分析模型在测试集上的表现，可以观察到预测值在测试集上的分布与真实值一致，且误差不随时间积累，说明模型具有较好的泛化能力。

4. 异常点与市场事件的对应分析：

检查回归图中预测误差较大的点可以发现与特定的市场事件或异常波动相关。例如，2024 年 9 月 24 日国新办举行新闻发布会，发布多项重磅政策、国际事件如巴以地缘风险，乌俄战争，美联储持续降息等可能会导致市场情绪波动，从而对股市价格产生较大影响。

而模型的误差在这些特殊时间段显著增大，说明模型在捕捉突发事件导致的市场非线性变化方面存在不足。

5. 对情感指标的解释与分析：

情感指标（BI_Simple）在模型中起到了重要作用。结合回归图，发现市场情绪波动较大的时候（如正面或负面情感急剧变化），模型的预测值与真实值之间的差距有所缩小，说明情感指标有效捕捉了投资者心理对股价的影响。

6. 实际应用中的参考意义：

短期投资决策：回归图显示模型能够在大多数情况下准确预测收盘价，说明该模型可以作为投资者短期决策的参考工具。例如，预测价格与实际市场价格的差值可以用于捕捉短期套利机会。

辅助风险控制：通过分析预测误差的时间分布，可以识别出模型不稳定的时间段（如高波动期），为投资者在这些时段规避风险提供参考。

3.2 不足

本文在计算情感指数的时候使用了十分简单的公式，包没有过多考虑其他元素，可能会影响情感指数的可信度，这些都会影响模型的精确度，同时情感词典并未含括所有的情感词汇，可能会影响一些句子的积极和消极属性。在实际操作中，评估人员需要收集大量的相关数据，多方面的考虑情感指数的计算变量和关系，由此得到的情感指数才更加接近于真实情况。本项目只是使用了较简单的基于支持向量机和 XGBoost 的方法，并未探讨其他方法所得出的模型的准确度。

成员工作总结与体会

杨穗瑜：

在本次大作业中，我负责统筹推进大作业的各个关键环节，既锻炼了我的领导能力，也加深了我对项目管理的理解。

首先，我负责了金融行业和科技行业股票评论数据的爬取工作，使用网页爬虫技术高效地获取了数万条股评数据。这是我第一次系统地进行大规模数据爬取，我通过实际操作提高了自己的数据抓取技巧，学会了如何应对网页结构的变化和数据清洗的挑战。

在模型搭建与优化方面，我负责了分类与回归模型的构建工作。通过选择合适的机器学习方法，并对模型进行调优，完成了股票评论的情感分类和上证指数走势的回归预测。在这个过程中，我深入学习了机器学习中的特征工程、模型评估以及调参技巧，不仅提升了我的模型构建能力，还加深了对算法背后原理的理解。

在此基础上，我进一步通过日期分组构建情感指数，以分析情感波动对股票市场表现的影响。同时，将每日的情感指数与上证指数相关数据合并，并进行可视化，直观地展示了情感波动与股票市场表现之间的联系。

通过本项目，我不仅在技术上取得了显著进步，还学会了如何在团队中协调工作、分配任务，提升了团队协作和沟通能力。此外，我还更加深入地理解了金融行业的实际应用，并积累了宝贵的实践经验，这对我未来的职业发展无疑是一次重要的提升。

孙玲：

在本次项目中，我承担的工作极大地锻炼了我的多方面能力，同时也让我对项目内容有了更为深入的理解。

首先，在数据获取环节，我负责能源和工业行业股票评论数据的爬取工作。我运用网页爬虫技术成功获取了大量的股评数据。这一过程并非一帆风顺，我需要应对不同网页结构带来的挑战，同时还要处理爬取后的数据清洗工作。但正是通过这样大规模的数据爬取实践，我的数据抓取能力得到了很大提升，学会了许多处理复杂网页结构和清洗数据的有效方法。

在数据分析与模型构建方面，我负责了不同类别股票情感和市场表现相关的工作，包括计算总的情感以及不同类别的股票和上证指数的相关系数。这要求我深入理解股票市场数据的特点，掌握数据分析的方法和技巧。通过这些工作，我学会了如何从海量的数据中挖掘有价值的信息，如何运用统计方法准确地计算相关系数，这使我对股票市场的内在联系有了更深刻的认识。

此外，我还负责各行业词云图的生成工作。词云图能够直观地反映各行业的热点话题和关键词分布，为进一步的分析提供了可视化的依据。在这个过程中，我熟练掌握了生成词云图的工具和技术，学会了如何根据数据特点选择合适的参数，以确保词云图能够准确、有效地传达信息。

通过本项目，我在技术能力上取得了长足的进步。不仅掌握了数据爬取、数据分析、词云图生成等技术，还深入学习了股票市场相关知识。同时，在团队协

作方面，我学会了如何与团队成员有效沟通、协调工作、合理分配任务，这大大提升了我的团队协作能力。这些收获对我的职业发展具有非常重要的意义，为我今后在相关领域的工作积累了宝贵的经验。

吴嘉文：

在项目中，我负责对情感指数进行计算的部分、以及对于数据的预处理部分、各行业消极和积极词汇统计的柱状图的生成工作、以及部分爬虫工作。

通过这一次数据挖掘大作业的实践，我收获了许多。从大作业选题、数据采集、特征选取、数据预处理，到模型训练、相关性分析、数据可视化，我们小组不断讨论，不断完善，一步一个脚印，一边学习理论知识，一边用软件与代码将理论转化为实践。探索的路上并非是一帆风顺的，在完成该项目的路上，我们也常常遇到问题。最初的数据集采集我们利用爬虫技术爬取相关的数据集，大家其实都不是很会写怕从，于是边学边写，过程遇到很多的问题，特别是反爬机制的问题。在这过程中既需要耐心学习，也需要耐心等待。数据预处理过程中，我们则将课本中所学的知识运用起来，将缺失值用均值填充，采用最大-最小值规范化进行特征缩放，真正了解了书本中一个方法背后的原理。这一次的合作，也培养了我的团队协作能力以及统筹规划能力，我非常感谢我的组员，能和他们一起齐心协力完成一件事，我感到很开心也很荣幸。

参考文献

[1] 洪巍，李敏. 文本情感分析方法研究综述[J]. 计算机工程与科学, 2019, 41(4): 750-758.

[2]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(8): 1769-1778.

[3] 许雪晨，田侃. 一种基于金融文本情感分析的股票指数预测新方法[J]. 数量经济技术经济研究, 2021(12): 9-22.

[4] Kumar, K. S., Radha Mani, A. S., Ananth Kumar, T., Jalili, A., Gheisari, M., Malik, Y., Chen, H.-C., & Moshayedi, A. J. (2024). Sentiment Analysis of Short Texts Using SVMs and VSMs-Based Multiclass Semantic Classification. Applied ArtificialIntelligence,38(1),e2321555.https://doi.org/10.1080/08839514.2024.2321555

[5] 李洋，董红斌. 基于CNN和BiLSTM网络特征融合的文本情感分析[J]. 计算机应用, 2018, 38(11): 3075-3080.

[6]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(4): 86-101.

[7] Rizinski, M., Peshov, H., Mishev, K., Jovanovic, M., & Trajanov, D. (2024). Sentiment Analysis in Finance: From Transformers Back to eXplainable Lexicons (XLex). IEEE Access, 12, 7170-7198.

[8] Peivandizadeh, A., Hatami, S., Nakhjavani, A., Khoshcima, L., Chalak Qazani, M. R., Haleem, M., & Alizadehsani, R. (2024). Stock Market Prediction With Transductive

Long Short-Term Memory and Social Media Sentiment Analysis. IEEE Access, 12, 87110-87130. https://doi.org/10.1109/ACCESS.2024.3399548