

融合网络新闻主题分布与主题分类下新闻情感的原油价格预测研究

余乐安¹, 赵晨珊², 宋正阳¹

(1. 四川大学商学院, 成都 610065; 2. 北京化工大学经济管理学院, 北京 100029)

摘要 石油在全球经济发展中扮演着重要的角色。为了更有效地预测石油价格, 本文将新闻媒体和舆论事件等信息纳入到油价预测研究中。首先, 应用主题识别模型对与油价相关的网络新闻进行主题识别, 将其转化为主题分布数据。并根据新闻主题对新闻进行情感分析, 最后将主题分布与每个主题下的新闻情感特征加入到油价预测模型中, 以此来改进预测效果。实证结果表明, 相较于基准模型, 结合了网络新闻主题特征以及主题分类下新闻情感特征的预测模型具有更好的预测性能。这表明, 新闻媒体中蕴含的信息能够有效反映市场情绪, 并有助于对油价进行预测。

关键词 网络新闻; 主题分布; 情感分析; 文本挖掘; 油价预测

Crude Oil Price Forecasting with Online News Topic Distribution and News Sentiment Classified by Topics

YU Lean¹, ZHAO Chenshan², SONG Zhengyang¹

(1. Business School, Sichuan University, Chengdu 610065, China; 2. School of Economics and Management, Beijing University of Chemical Technology, Beijing 100029, China)

Abstract Crude oil plays an important role in global economic development. To effectively predict crude oil prices, this paper incorporates news media into oil price forecasting. First, we apply a topic recognition model to identify the topics of online news and transform them into topic distribution data. Then we conduct sentiment analysis on news articles based on their topics, finally, we integrate the topic distribution and news sentiment features under each topic into the oil price prediction model to enhance its accuracy. The empirical results indicate that integrate network

收稿日期: 2023-01-17

基金项目: 国家自然科学基金重点项目 (72034003)

Supported by Key Program of National Natural Science Foundation of China (72034003)

作者简介: 通信作者: 余乐安, 教授, 博士生导师, 研究方向: 商务智能、能源预测, E-mail: yulean@amss.ac.cn; 赵晨珊, 硕士, 研究方向: 能源预测, E-mail: z_shan11@126.com; 宋正阳, 博士研究生, 研究方向: 金融预测、投资组合, E-mail: Szy_atuo@163.com.

news topic features and news sentiment features under topic classification performs better than the benchmark model. This suggests that the information contained in news media can effectively reflect market sentiment and contribute to the forecasting of crude oil prices.

Keywords online news; topic distribution; sentiment analysis; text mining; crude oil price forecasting

1 引言

原油是全球金融市场上最重要的大宗商品之一,其价格波动往往从多个方面对世界宏观经济产生重大影响。因此,预测原油价格并且提高预测精度已成为该领域关键的热点问题。然而,许多复杂的驱动因素使得原油价格难以预测,以往很多预测模型经常产生较大的预测误差。这就需要探索新的预测建模思路来减少预测误差。

实际上,原油价格预测研究已经持续了很多年,一直是学界研究的热点之一。根据现有相关文献,预测方法主要分为两大类:传统计量模型和人工智能算法模型。基于传统计量经济模型的预测方法应用较早,研究成果也相当丰富。例如, Murat and Tokat (2009) 使用随机游走 (RW) 模型作为基准模型来预测石油市场的价格趋势并评估预测性能。 Coppla (2008) 应用向量误差修正模型 (VECM) 预测原油期货价格的波动。 Morana (2001) 提出一种基于广义自回归条件异方差 (GARCH) 模型的半参数方法来预测布伦特原油的短期价格。 Baumeister and Kilian (2012) 利用向量自回归 (VAR) 模型短期预测油价,实证结果表明, VAR 模型比自回归 (AR) 模型和自回归移动平均 (ARMA) 模型具有更高的方向预测精度。我国学者也有大量的研究成果,侯璐 (2009) 以布伦特原油的现货报价为依据,建立 ARIMA 预测模型,并且分析了 2009 年国际以及石油行业的新局势和动态变化。赵沙 (2013) 按照时间序列分析的方法构建了指数广义自回归条件异方差模型 (EGARCH)、季节-谐波预测模型以及指数平滑模型,并将这三种模型用于国际油价预测中。通过对预测效果的对比得出结论:季节-谐波预测模型相对较好。

传统的计量经济模型具有处理线性数据的基本假设,因此限制了它们的应用范围。随着人工智能的发展,人工智能算法能进行自组织自适应学习,而且可以不断调整参数,对非线性数据的处理显现出明显的优势。因此,人工智能算法受到越来越多学者们的青睐,在预测领域也取得了许多重要的成果。例如, Mirmirani and Li (2004) 使用基于遗传算法的人工神经网络模型来预测油价,预测结果表明,该模型显著优于向量自回归 (VAR) 模型。 Kaboudan (2001) 将遗传算法和人工神经网络算法在月度油价短期预测中的预测性能进行了比较与评价。 Urolagin et al. (2021) 使用多变量长短期记忆模型 (LSTM) 进行油价预测。 Lu et al. (2022) 对比了大量机器学习以及神经网络模型和传统的预测模型在预测原油期货上的表现,其实证结果表明相比于传统的预测模型,机器学习模型有更好的表现。 Yu et al. (2015) 利用极限学习机 (ELM) 和扩展极限学习机 (EELM) 模型对 WTI 原油价格进行预测。 Tang et al. (2018) 使用随机向量函数链路 (RVFL) 预测 WTI 油价,预测结果表明,没有迭代训练过程的 RVFL 计算时间短,预测精度更高。国内学者也有大量的研究成果,郑俊艳 (2012) 构建了一种基于小波分析和支持向量机的原油价格预测模型,实现了对 WTI 国际原油价格走势

的预测. 赵阳 (2015) 提出基于压缩感知去噪的人工智能预测方法来预测油价, 该模型表现出较好的预测效果与稳定性. 这些人工智能预测模型都大幅提升了油价预测的效果.

然而, 无论是传统经济计量预测模型还是人工智能预测模型, 建模都是基于数值型的数据, 对非数值型数据关注不够. 近年来, 在对原油价格的预测研究中, 石油相关事件的影响引起了许多学者的关注. 根据以往的相关研究, 原油价格的大幅波动主要是由石油相关事件引起的, 这些事件可以看作是对整个石油行业的冲击. 例如, Zhang et al. (2009) 对 1991 波斯湾战争和 2003 伊拉克战争对于油价的影响进行了实证研究. 实证结果表明, 事件对原油价格波动产生影响, 并且对原油市场产生巨大的影响. Ji and Guo (2015) 研究了四种与石油有关的事件 (飓风、全球金融危机、战争和欧佩克会议) 对世界油价的影响, 并确定了市场对石油相关事件的反应程度和重要性. 国际油价的长期走势最终取决于原油的供应和需求. 然而, 石油相关事件通过影响原油供需状况以及影响市场参与者的心理预期, 会放大油价波动, 进而增加原油市场的不稳定性. 简而言之, 在进行油价预测时, 考虑石油相关事件的影响非常重要, 而这在传统的预测模型中经常被忽略.

传统的数据源具有时滞性, 无法捕捉到石油相关事件的即时影响, 然而, 随着互联网技术和大数据技术的发展, 实时获取丰富的网络信息成为可能. 近年来, 越来越多的学者利用网络信息进行相关领域的研究, 并取得了丰硕的研究成果. 通过网络文本信息, 例如网络新闻、推特 (Twitter)、论坛等, 可以即时获取与石油相关的事件信息. 其中, 网络新闻是更好的研究选择, 因为网络新闻的写作方式相较于其他网络文本来说更加规范化, 更大程度上减少了噪音, 而且权威新闻网站提供的新闻是由领域内专业的经济、金融记者或相关专家撰写, 他们对新闻事件发表的观点具有更广泛的影响力. 更重要的是, 网络新闻的预测能力已经被证实. 例如, Wex et al. (2013) 用回归分析的方法从统计学上证明了网络新闻的预测能力, 在此基础上, 构建了全球石油危机预警系统. 此外, 对网络新闻进行情感分析是量化相关事件影响的一种方式.

情感分析 (又称意见挖掘) 是理解文本信息的一种有效的数据挖掘工具. 通常情况下, 可以通过情感分析在文本文件中提取主观情绪, 进而衡量文本的内容是积极的还是消极的. 例如, Skuza and Romanowski (2015) 利用机器学习对来自推特上的大量文本数据进行情感分析, 进而对未来的股价做出估计, 证明了情感与股票市场之间的相关性并且证实了所选方法的有效性. Smales and Lee (2016) 研究了新闻情感、股市标准普尔 500 指数 (SPX) 收益与隐含波动率指数 (VIX) 变化之间的关系. 研究结果显示, 波动率指数的变化与新闻情感以及股票收益之间存在显著的负相关关系. 当期和滞后的新闻情感与股票收益呈显著的正相关关系, 而且新闻情感能够预测未来的股票收益. 同时, 许多国内学者对于情感分析用于预测领域也有大量研究, 陈晓红等 (2016) 对新浪微博文本应用情感分析技术, 构建情绪指标, 考察投资者情绪对于股票市场的预测能力. 研究结果表明, 微博情绪指数对于股价的预测作用在短期内非常显著. 陈海文等 (2016) 将黄金收盘价历史价格数据与挖掘财经新闻中的情感倾向信息相结合, 构建出一种基于财经新闻信息挖掘的金融品种价格走势预测方法, 该方法在实证中显现出较好的预测效果. Kraaijeveld et al. (2020) 使用格兰杰因果检验、基于加密货币特定词汇的情绪分析方法探究了 Twitter 上的公众情绪对比特币等加密货币价格回报的预测能力, 其结果表明 Twitter 情绪对比特币、莱特币等具有较强的预测能力. Naseer et

al. (2022) 研究了加密货币走势与投资者情绪之间的相关性, 并提出了一种结合深度学习算法和卷积神经网络的方法对加密货币进行预测, 结果表明投资者情绪能够提高预测的准确率. Aharon et al. (2022) 的研究结果表明社交媒体中表达的投资者情绪与资产价格之间存在着强烈的因果关系.

在金融市场领域之外, 情感分析也已经应用于油价研究领域. 例如, Alfano et al. (2015) 根据噪声交易者方法, 考察油价相关新闻情感对不同类型投资者的影响. 其实证结果表明, 新闻情感不仅对噪声残差有显著的正向影响, 而且对资产的基础价格也有显著的正向影响. Lechthaler and Leinert (2012) 研究了 2003 年后全球原油市场的价格动态走势. 作者通过对汤森路透社新闻分析数据库中的新闻进行分析, 并将其处理为情感时间序列, 来解释结构性中断和估计市场预期. 研究表明, 新闻情感对油价变化具有显著的正向影响. Li et al. (2016) 对汤森路透社新闻进行情感分析, 将新闻情感作为一个重要的变量加入 WTI 原油价格预测模型中, 成功地提高了国际油价预测性能.

本文主要创新在于: 过往研究较少关注网络新闻之间的主题关联, 而主题分布的变化间接表示石油相关事件可能所属类别的变化情况, 同时经过主题划分的新闻篇章情感强调主题聚合下的情感, 每一个主题下的情感都可视为一个影响因素. 因此, 基于网络新闻的主题分布以及每个主题下的新闻情感, 本文提出了一个融合网络新闻主题分布与主题分类下新闻情感的油价预测方法, 通过主题识别模型和情感分析技术, 对包含大量有关事件和舆论因素的网络新闻进行主题识别, 将其转化为主题分布数据, 并通过主题对新闻情感进行分类, 之后将主题分布与每个主题下的新闻情感分别加入到原油价格预测模型中. 基于此, 本文运用主题模型对网络新闻隐含主题进行建模, 找到众多网络新闻之间的关联, 得到主题分布, 并按照主题对新闻情感进行分类, 之后利用每个交易日的新闻主题分布以及主题分类下的新闻情感来预测油价. 因此, 本文将聚焦融合网络新闻主题分布及主题分类下新闻情感对油价预测的影响, 重点分析对油价预测效果的提升.

2 模型构建

2.1 模型框架

本文的研究方法是对包含大量有价值信息的石油相关网络新闻进行主题识别, 将其转化为主题分布数据, 并通过主题对新闻情感进行分类, 将主题分布与每个主题下的新闻情感分别加入到原油价格预测模型中, 以此来改进油价预测模型的预测效果. 该方法可以称为融合网络新闻主题分布和主题分类下新闻情感的原油价格预测方法, 其一般框架描述见图 1 所示. 不难看出, 该方法主要包括四个步骤: 文本获取及预处理、主题识别与主题分布计算、情感分析、原油价格预测, 下面对框架模型中的四个主要步骤进行简要概述.

步骤 1: 文本获取及预处理

收集和预处理油价相关网络新闻是该方法的首要工作. 首先, 基于文本收集或网络爬虫技术, 获取网络新闻和新闻发布的时间. 之后, 将同一天发布的所有新闻分成一组, 由于这里只研究交易日的情况, 需要将非交易日的新闻从新闻数据集中过滤出去. 最后, 需要对新闻进行预处理, 如检查和消除重复记录和缺失记录, 对文本进行小写化、分词、去除停用词等, 从而得到较为“干净”的新闻文本数据.

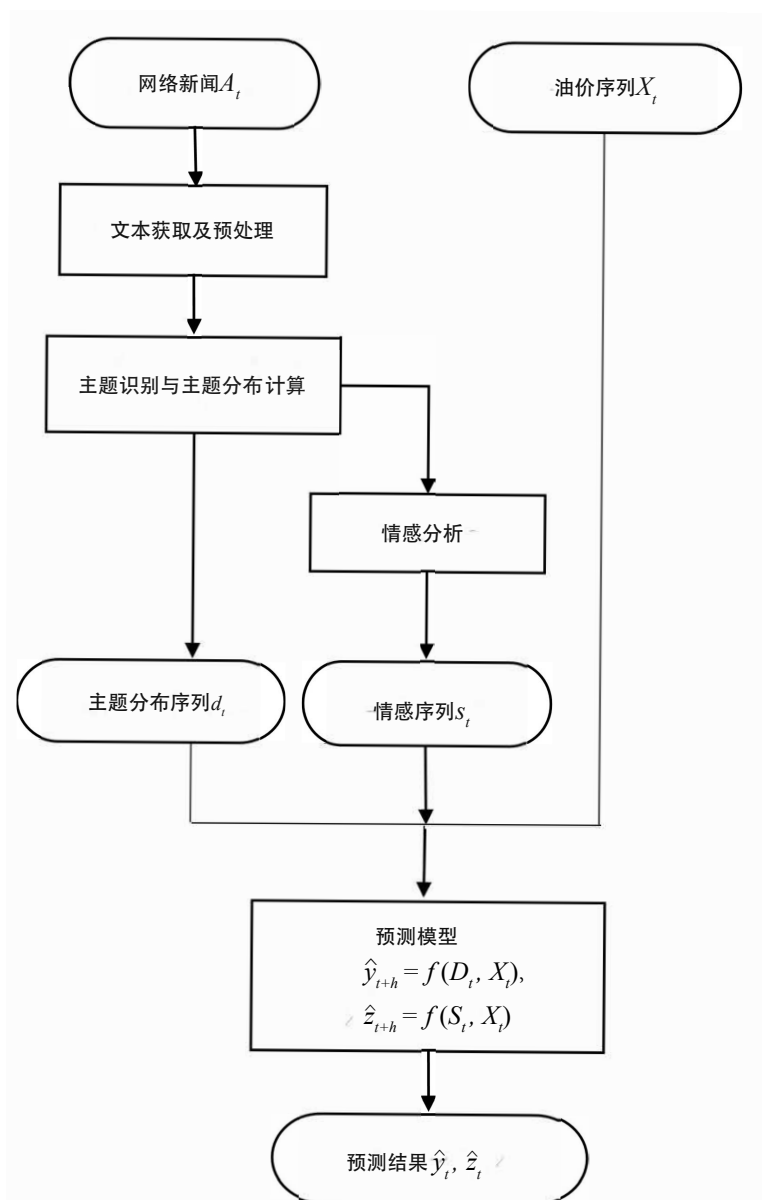


图 1 融合网络新闻主题分布与主题分类下新闻情感的原油价格预测方法框架

步骤 2: 主题识别与主题分布计算

对于第一个步骤收集的网络新闻文本数据, 采用主题模型来识别网络油价相关新闻的主题, 并通过计算得到相应主题分布。这里主要选择了三种经典的主题模型来识别新闻主题, 即潜在狄利克雷分布模型 (latent Dirichlet allocation, LDA)、概率潜在语义分析模型 (probabilistic latent semantic analysis, pLSA) 和分层狄利克雷过程模型 (hierarchical Dirichlet processes, HDP), 其具体描述见 2.3 节。

步骤 3: 情感分析

这个步骤主要是将网络新闻按照概率最大的主题进行分类. 为了从油价相关网络新闻中提取有价值的信息, 运用情感分析技术, 对每一个主题分类下的新闻进行情感分析, 从而得到相应的情感极性分值.

步骤 4: 原油价格预测

将网络新闻主题分布和每一个对应主题的新闻情感分值分别加入原油价格预测模型中, 以此来改进油价预测模型, 并对预测效果的改善进行分析.

2.2 情感分析

在情感分析部分, 本文使用了 Python 的一个库 Textblob, 它提供了一种较为方便的方法来深入研究常见的自然语言处理 (natural language processing, NLP) 任务, 比如情感分析. 在现有文献中, Li, Shang and Wang (2017) 使用 Textblob 进行情感分析进而从石油相关新闻标题中提取情感特征, 并将情感特征加入到油价走势和波动预测中, 取得较好的预测效果. 下面对 Textblob 的情感分析功能做一个简要介绍.

Textblob 的情感分析模块返回每个文档的情感极性分数. 极性分数是 $[-1.0, 1.0]$ 范围内的值, 其中小于 0 为负向情感, 大于 0 为正向情感. Textblob 的情感分析是基于语义的. 其情感词典里的词是按照词义做标注的, 而且每个词的不同语义都有对应的极性、主观性和强度分数. Textblob 不是简单地识别并计算正向词和负向词的数量, 而是根据上下文语境不断调整情感极性, 得到最终的情感极性分数. 下面以一个具体的例子来说明.

在 Textblob 的情感词典中, “excellent” 这个单词的情感极性分数为 1.0, “very” 的情感强度分数为 1.3, 对于 “very excellent” 这个短语, TextBlob 将 “very” 识别为修饰词, 会忽略它的极性, 只使用强度来修正后面的单词, 那么情感极性分数为 1 乘以 1.3 等于 1.3. 但由于这个情感极性分数超过了 1.0, 而情感极性分数是一个在 -1.0 至 1.0 的区间数, 所以最终 “very excellent” 的情感极性分数是 1.0. 此外, 否定词的存在使极性乘以 -0.5 , 因此 “not excellent” 短语的极性是 -0.5 . TextBlob 在文本中继续寻找可以分配极性的单词和短语, 并将它们平均用于较长的文本. 依照这样的方法, 可以对新闻篇章的文本进行情感分析, 得到相应的新闻情感分值.

2.3 主题模型

主题模型指的是对文本中隐含主题进行建模, 找到文本间关联的方法, 这里选择三种经典的主题模型来识别新闻主题, 即 LDA、pLSA 和 HDP 模型. 详细介绍如下.

1) 潜在狄利克雷分布模型 (latent Dirichlet allocation, LDA)

LDA 主题模型由 Blei, Ng and Jordan (2003) 于 2003 年提出, 它是一种文档生成模型, 使用词袋模型 (bag-of-words model) 即不考虑词的顺序. LDA 是无监督机器学习模型的典型应用之一, 具有快速、高效的优点, 因此被众多学者广泛研究与应用.

在 LDA 模型中, 各个主题在文档中出现的概率即主题分布与各个词在某主题下出现的概率分布即词分布是不确定的. LDA 采用非参数贝叶斯模型的思想, 认为主题分布与词分布应该服从两组多项式分布, 因为多项式分布和狄利克雷分布是共轭结构, 所以狄利克雷分布在 LDA 主题分布和词分布中被用作为它们的共轭先验分布. 通常生成文档方式如图 2 所示.

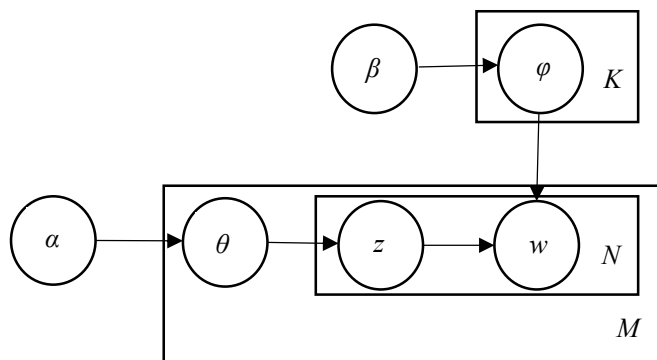


图2 LDA 的概率图模型

在图2中, M 表示文档的总数, N 为给定文档的单词总数, K 为主题个数, w 是能被观察到的变量, 除此之外是参数或者隐变量. 图中的 θ 为主题分布, φ 为词分布, α 、 β 为狄利克雷分布参数, α 作为主题分布 θ 的先验分布, β 则是作为词分布 φ 的先验分布. 参考 Griffiths and Steyvers (2004), 一篇文档生成的具体过程描述如下:

- ① 根据文档先验概率 $p(d_i)$, 选一篇文档 d_i ;
- ② 根据选择的文档 d_i , 由狄利克雷分布 α 里取样生成 d_i 的主题分布为 θ_i ;
- ③ 根据主题分布 θ_i , 从该分布里取样生成主题 $z_{i,j}$, 即为文档 d_i 的第 j 个词;
- ④ 由另一个狄利克雷分布 β 里取样生成主题 $z_{i,j}$ 对应的词语分布 $\phi_{z_{i,j}}$;
- ⑤ 根据词语分布 $\phi_{z_{i,j}}$, 采样生成词语 $w_{i,j}$;
- ⑥ 重复 ① 至 ⑤ 这个过程 N 次, 可得到一篇文档.

对于参数 θ 和 φ , 使用吉布斯采样方法 (Gibbs sampling) 进行采样是最广泛的做法, 具体 Heinrich (2005) 在文献中有详细描述. 最终得到参数的估计为:

$$\hat{\theta}_{mk} = \frac{n_{m,\neg i}^{(k)} + \alpha_k}{\sum_{k=1}^K (n_{m,\neg i}^{(k)} + \alpha_k)}, \quad (1)$$

$$\hat{\varphi}_{kt} = \frac{n_{k,\neg i}^{(t)} + \beta_t}{\sum_{t=1}^V (n_{k,\neg i}^{(t)} + \beta_t)}, \quad (2)$$

其中 “ $\neg$ ” 的意思是将当前的词汇去除掉, $n_{k,\neg i}^{(t)}$ 为分配于主题 k 中词 t 的次数 (去掉当下分配的词 i), $n_{m,\neg i}^{(t)}$ 为分配于主题 k 且在文档 m 中的词的总计个数 (去掉当下分配的词 i). 通过上述处理, 最终可得到如下公式:

2) 概率潜在语义分析模型 (probabilistic latent semantic analysis, pLSA)

pLSA 主题模型是由 Hoffman 于 1999 年提出的一种较为早期的主题模型. 该模型也采用词袋模型, 生成文档的方式如图3所示, 具体步骤描述如下:

- ① 按照文档先验概率 $p(d_i)$, 选择一篇文档 d_i ;
- ② 根据选择的文档 d_i , 从主题分布中按照概率 $p(z_k | d_i)$ 选择一个隐含的主题类别 z_k ;
- ③ 根据选择的主题 z_k , 从词分布中按照概率 $p(w_j | z_k)$ 选择一个词 w_j ;

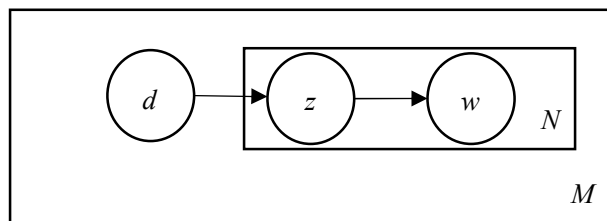


图3 pLSA 的概率图模型

④ 重复 ① 至 ③ 这个过程 N 次, 可得到一篇文档。

根据 pLSA 生成文档的过程与 LDA 对比可知, LDA 是在 pLSA 的基础上, 为各个主题在文档中出现的概率分布 (即主题分布) 与各个词语在某个主题下出现的概率分布 (即词分布) 分别加了两个狄利克雷先验。pLSA 的主题分布与词分布都是唯一确定的, 但 LDA 不同, 在 LDA 中将它们认为是随机变量。

在 pLSA 中, 可以观察到的数据是 $(d_i; w_j)$ 数据对, 而 z_k 是隐含变量。文档中每个词的生成概率为:

$$p(d_i, w_j) = p(d_i) \sum_{k=1}^K P(w_j | z_k) P(z_k | d_i), \quad (3)$$

其中 $P(w_j | z_k)$ 、 $P(z_k | d_i)$ 分别服从两组多项式分布, 需要估计这两组分布的参数, 一般使用 EM 算法 (最大期望算法)。EM 算法的步骤简要介绍如下:

E 步骤: 假设已知 $P(w_j | z_k)$, 求隐含变量 $P(z_k | d_i)$ 在当前估计参数条件下的后验概率 $P(z_k | d_i, w_j)$;

M 步骤: 对参数 $P(w_j | z_k)$ 、 $P(z_k | d_i)$ 的最大化全变量对数似然函数的期望求最大值, 此时使用 E 步骤里计算的隐含变量后验概率, 得到新的参数值。

两步迭代循环计算直到收敛。具体计算过程可参考 Hoffman (1999) 的论文。

3) 分层狄利克雷过程模型 (hierarchical Dirichlet processes, HDP)

HDP 主题分布模型由 Teh et al. (2006) 于 2006 年提出, 是一种分层聚类的模型。该模型可以自动确定聚类的数量, 能针对多组数据进行建模。在 HDP 模型中, 假设存在多组数据, 每一组数据都存在未知数量的类别。HDP 在计算过程中能够找出每一组数据中包含的聚类结果, 因此它与 LDA 不同, 不用对主题的个数进行选择, 该模型可以估计出恰当的主题个数。生成文档的方式如图 4 所示。

结合图 4 基于 HDP 的文档生成过程可描述如下: 假设所有文档的主题分布都来自于某个分布 H , 那么可将 λ 和 H 作为狄利克雷分布的先验参数。在生成某篇文档的过程中, 起初由该先验分布中抽取一个分布 G_0 , 作为这篇文档主题的先验分布, 再以 G_0 和 α_0 为参数构造一个狄利克雷分布, 然后抽取一个主题分布 G_j 作为第 j 篇文档的主题分布, 并从该主题分布中抽取第 i 个单词的主题 θ_{ji} , 最后从该主题中抽取一个单词 x_{ji} 。因为所有文档都共享

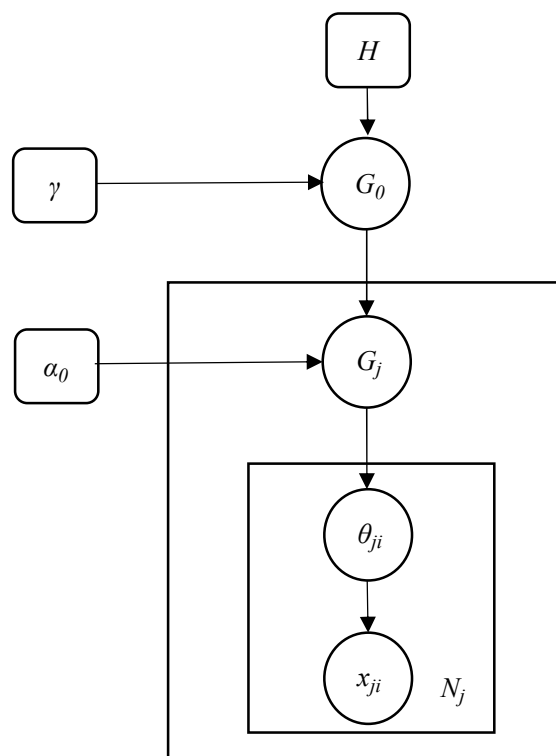


图4 HDP 的概率图模型

一个主题先验分布 G_0 的参数 H , 而每个文档主题先验分布 G_0 都是从一个以 H 为基分布的 DP 中抽取, 因此所有的文档都共享了相同的主题先验分布。

HDP 模型的参数存在多种构造方式, 如折棍构造 (Stick-breaking) 与中国餐馆连锁构造 (Chinese restaurant franchise, CRF) (Brey (1996)), 都能体现良好的聚类性质, 本文使用的是 CRF (中国餐馆连锁) 构造的模型。CRF 构造的核心思想为: 假设有一个餐厅特许经营权 (语料库), 连锁经营的有许多餐厅 (文档, j), 该连锁下的所有餐厅共用同一个菜肴菜单 (主题, ϕ_k)。而且, 一个菜肴 (主题, ϕ_k) 只放在单张桌子 t 上, 专门为坐在那张桌子上的所有顾客 (单词, $\theta_{j,i}$) 所提供。此外, 当顾客进入该餐厅时, 可以自主选择坐在任何想坐的地方, 该客人可以选择坐在已经有一些客人坐过的桌子, 也可以选择坐没人坐过的新的桌子。这里选择每个选项的概率是相同的。现在, 连锁菜单中的菜肴对应于全局参数主题 ϕ_k , 每个餐厅对应于一个单独的文档 j 。因此, 特定餐厅中提供的菜肴数量对应于特定文档中的主题个数, 每张桌子上坐着的人数对应着 j 文档中每个主题的字数。采用 CRF 中的概念很容易转移到 HDP 的构造中。

2.4 原油价格预测

现有文献中, 广泛运用的原油价格预测方法是一个典型的没有外部因素的时间序列预测方法, 即基于历史观测值 $X_t = \{x_t, x_{t-1}, \dots, x_{t-(m-1)}\}$, 计算得到原油价格在步长为 h 时的

预测值 $\hat{y}_{t+h}$, 如下式所示:

$$\hat{y}_{t+h} = f(X_t) = f(x_t, x_{t-1}, \dots, x_{t-(m-1)}), \quad (4)$$

其中, m 是自回归的滞后阶数, f 为预测模型形式. 在油价预测中, 如果将新闻主题分布 D_t 加入到油价预测方法中, 公式 (4) 可以推广到公式 (5):

$$\hat{y}_{t+h} = f(D_t, X_t) = f(d_{t-l+1}, \dots, d_t, x_{t-(m-1)}, \dots, x_t). \quad (5)$$

此外, 将新闻主题分类下的情感分值 S_t 加入到油价预测方法中, 公式 (4) 可以推广到公式 (6):

$$\hat{y}_{t+h} = f(S_t, X_t) = f(s_{t-l+1}, \dots, s_t, x_{t-(m-1)}, \dots, x_t), \quad (6)$$

其中, l 是情感的预测滞后期.

在上面 3 个公式的预测模型形式选择中, 为了对比, 本文不仅采用了线性模型: 线性回归模型 (linear regression, LR) (Heinrich (2005)), 而且采用三个非线性的人工智能模型: 随机向量函数链路网络模型 (random vector function link, RVFL) (Pao et al. (1994) and Zhang and Suganthan (2016)), 支持向量回归模型 (support vector regression (Cortes and Vapnik (1995) and Burges (1998)), SVR) 和极限学习机模型 (extreme learning machine, ELM) (Huang et al. (2006)) 作为预测模型, 以此来研究网络新闻主题分布和主题分类下新闻情感对油价的预测能力. 三个非线性预测模型的介绍如下:

1) 随机向量函数链路网络模型 (RVFL)

随机向量函数链路网络模型 (RVFL) 是一个单隐层前馈网络, 它具有良好的非线性拟合能力. 因此, 本文采用 RVFL 作为 WTI 原油价格的预测模型, RVFL 的核心思想是利用隐含层学习的增强非线性核原始数据来提高泛化能力, RVFL 模型的输出可表示为:

$$\hat{y}_i = \sum_L^{j=1} \beta_j h(\omega_j x_i + b_j) + \sum_{L+n}^{j=L+1} \beta_j x_{ij}, \quad i = 1, 2, \dots, N, \quad (7)$$

其中 $x_i = [x_{i1}, x_{i2}, \dots, x_{in}]$ 是 n 维向量, n 代表输入的特征维度, $\hat{y}_i$ 是预测值, L 表示隐藏层的节点数量. w_j 和 b_j 分别代表隐藏层的输入权重和权重. 两个参数 β_j 代表输出权重, 函数 $h(x)$ 是激活函数, 通常使用 Sigmoid 函数作为激活函数. 隐藏层的输入矩阵通常可以写成下面的形式:

$$H = \begin{bmatrix} h(\omega_1 x_1 + b_1) & \cdots & h(\omega_L x_1 + b_L) & x_{11} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ h(\omega_1 x_N + b_1) & \cdots & h(\omega_L x_N + b_L) & x_{N1} & \cdots & x_{Nn} \end{bmatrix}, \quad (8)$$

$$\beta = \begin{bmatrix} \beta_1^T \\ \vdots \\ \beta_L^T \\ \vdots \\ \beta_{L+n}^T \end{bmatrix}_{(L+n) \times m}, \quad \hat{y} = \begin{bmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_m \end{bmatrix}_{m \times N}. \quad (9)$$

2) 支持向量回归模型 (SVR)

支持向量回归 (SVR) 基于支持向量机 (support vector machine, SVM) 算法, SVR 的核心思想是将回归问题转化为一个最优化问题, 通过寻找最优化的超平面来拟合数据. 与 SVM 相似, SVR 也将数据点映射到高维空间, 通过寻找一个最优的超平面, 使得距离该超平面最近的数据点与该超平面之间的距离最大化. SVR 根本上是把回归看作一个二次最优规划问题, 其凸二次性特征以及线性限制条件保证了其总是能获得全局惟一最优预测, SVR 的主要优点是能够处理高维数据和非线性数据, 具有较好的泛化性能和鲁棒性. SVR 的原问题表示为:

$$\begin{aligned} \min \quad & \frac{1}{2} \|w\|^2 + \frac{C}{n} \sum_{i=1}^n (\xi_i + \xi'_i), \\ \text{s.t.} \quad & w^\top \varphi(x) + b - y_i \leq \varepsilon + \xi_i, i = 1, 2, \dots, n, \\ & y_i - w^\top \varphi(x) - b \leq \varepsilon + \xi'_i, i = 1, 2, \dots, n, \\ & \xi_i \geq 0, \xi'_i \geq 0, i = 1, 2, \dots, n, \end{aligned} \quad (10)$$

其中 $x \in R^m$, $y_i \in R^1$, $\varphi(x)$ 为 SVR 的核函数, b 代表偏差, w 代表从特征层到输出层的线性权重, ξ_i 和 ξ'_i 是松弛变量. 通过构建拉格朗日方程可以得到 SVR 的最终回归决策方程:

$$f(x) = \sum_{i=1}^n w_i \varphi_i(x) + b. \quad (11)$$

3) 极限学习机模型 (ELM)

与 RVFL 相似, 极限学习机模型 (ELM) 同样是一种单隐层的前馈神经网络, 具有良好的非线性拟合能力, 并且具有较快的运算速度和泛化能力, 常用于时间序列的预测. ELM 模型的输出可表示为:

$$\hat{y}_i = \sum_{j=1}^L \beta_j h_j(x_i), \quad i = 1, 2, \dots, N, \quad (12)$$

其中 β_j 代表输出权重, 函数 $h(x)$ 是激活函数, ELM 隐含层的输出矩阵为:

$$\begin{bmatrix} h_1(x_1) & \cdots & h_L(x_1) \\ \vdots & \ddots & \vdots \\ h_1(x_N) & \cdots & h_L(x_N) \end{bmatrix}. \quad (13)$$

根据 Moore-Penrose 广义逆的方法, 可以求得 ELM 的输出矩阵:

$$\beta = H^+ Y. \quad (14)$$

3 实证分析

3.1 实验设计

本文使用的研究样本包括 WTI 原油价格日度数据与网络新闻文章日度数据, 其中 WTI 原油价格来源为美国能源信息署 (<http://www.eia.doe.gov/>), 网络新闻文章从“商业和金融新闻, 美国和国际即时新闻, 汤森路透社 (Business & Financial News, Breaking US & International News of Thomas Reuters)” (www.reuters.com) 检索得到. 选择路透社新闻有两个原因. 首先, 路透社提供了有关大宗商品市场大量有价值的消息, 且报道非常及时. 通过情感分析, 可以捕捉并量化各种交互因素对石油市场的影响. 其次, 路透社是一个权威的网络新闻网站, 其中新闻是由专业的经济、金融记者或专家撰写而成, 他们对新闻事件的看法具有更为广泛的影响. 本文研究样本数据区间为 2017 年 5 月 15 日至 2017 年 11 月 30 日, 共计 139 个交易日. 经过预处理, 共有 6647 篇新闻文章, 将它们映射到带有时间戳的数据并存储于数据库中. 同时, 将日度样本数据进一步分为训练集和测试集, 80% 为训练集, 20% 为测试集. 这样, 2017 年 5 月 15 日到 2017 年 10 月 20 日的数据作为训练数据, 共 111 个观测值, 占全部样本的 80%; 而测试数据是从 2017 年 10 月 21 日到 2017 年 11 月 30 日, 共 28 个观测值, 占全部样本的 20%.

对于预测模型, 这里采用典型的线性统计时间序列模型 LR 与三个人工智能模型: SVR、RVFL 和 ELM. 对于评价标准, 本文采用均方根误差 (root mean squared error, RMSE)、平均绝对百分比误差 (mean absolute percentage error, MAPE), 具体公式如下:

$$\text{RMSE} = \sqrt{\frac{1}{Q} \sum_{t=1}^Q (x_t - \hat{x}_t)^2}, \quad (15)$$

$$\text{MAPE} = \frac{1}{Q} \sum_{t=1}^Q \left| \frac{x_t - \hat{x}_t}{x_t} \right|, \quad (16)$$

其中, Q 是测试数据集的大小, x_t 和 $\hat{x}_t$ 分别是在时间 t 原油价格的实际值和预测值. 对于预测改进率度量指标 (improvement ratio, IR), 主要用来衡量测试模型相对于基准模型的优越性, 具体公式如下:

$$\text{IR}_{\text{MAPE}} = -\frac{\text{MAPE}_A - \text{MAPE}_B}{\text{MAPE}_B} * 100\%, \quad (17)$$

$$\text{IR}_{\text{RMSE}} = -\frac{\text{RMSE}_A - \text{RMSE}_B}{\text{RMSE}_B} * 100\%, \quad (18)$$

其中, 以 IR_{MAPE} 和 IR_{RMSE} 表示测试模型 A 相对于基准模型 B 的改进率. 显然, 如果 IR 为正, 可以证明目标模型具有较好的水平预测精度.

3.2 结果分析

根据前面的方法与实验设计, 首先要对油价相关的新闻文章进行预处理, 然后进行主题识别与主题分布计算, 最后进行情感分析以及原油价格预测. 其中, 3.2.1 节描述主题分布计算与主题分类下新闻情感的计算, 3.2.2 节主要讨论加入新闻主题分布以及主题分类下新闻情感后的油价预测结果.

3.2.1 主题分布计算与主题分类下新闻情感的计算

根据实验设计, 在对油价相关的新闻做好预处理后, 主要运用三种主题模型 LDA、pLSA 以及 HDP 来识别主题并计算主题分布, 从而得到划分的主题与对应主题的分布. 在此基础上, 计算主题分类下的新闻情感分值.

对于 LDA 主题模型, 首先需要确定合适的主题个数后才能进行后续分析. 这里用 Kullback-Leibler 散度 (Kullback-Leibler divergence, KL 散度, 也称相对熵) 来确定新闻主题个数. KL 散度, 是用来衡量两个概率分布之间差异的非对称度量. 主题分布之间的 KL 散度值越大, 则说明主题之间的差异越大, 主题划分的效果可能更好. 根据相关数据, 对不同主题个数下主题分布平均 KL 散度的计算结果如图 5 所示.

从图 5 可以看出, 主题个数为 13 时 KL 散度最大, 因此本文选择 13 作为主题的个数. 运用 LDA 主题模型, 很容易得到 13 个主题及其主题分布, 表 1 列出了这 13 个主题中概率最大的前 10 个词, 具体如表 1 所示.

从表 1 的结果可以看出每一个主题大致相关的内容, 比如 Topic1 是与欧洲相关, Topic3 是与 OPEC 供应相关, Topic11 是与股市情况相关的新闻. 充分说明用 LDA 主题模型提取的这些主题信息很大程度上反映了传统统计指标不能反映的信息, 比如 OPEC 政策的改变、不同地区的需求情况的变化以及商品市场的情况等大量对油价产生影响的多方面有价值的信息.

之后, 运用 pLSA 主题模型对油价相关新闻进行主题识别并得到主题分布, 各个主题中概率最大的前 10 个词具体如表 2 所示.

由表 2 也可以大致看出每一个主题相关的内容, 如 Topic8 是与 OPEC 供应相关, Topic10 可能与天然气相关, 但总体而言, 仅凭前十个词不足以明显判断出主题的确切内容.

最后, 采用 HDP 主题模型对油价相关新闻进行主题识别并得到主题分布, 各个主题中概率最大的前 10 个词具体如表 3 所示.

由表 3 同样能大致看出每一个主题相关的内容, 如 Topic5 是与印度相关, Topic10 可能

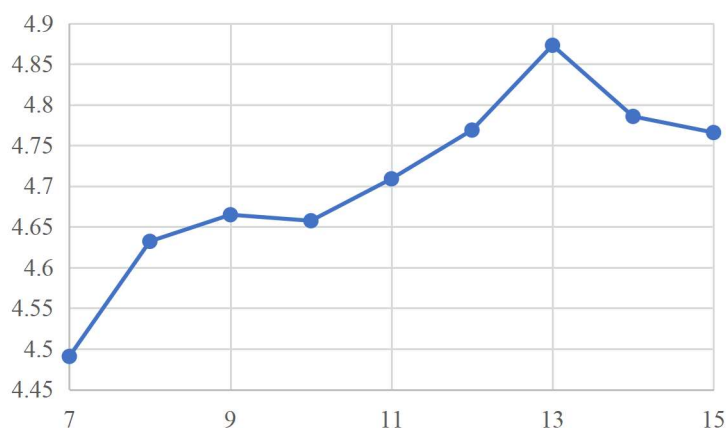


图 5 不同主题个数的平均 KL 散度

表 1 LDA 主题模型得到的相应主题下概率最高的十个词

主题	词
Topic0	rating fitch security issuer bank risk term debt credit party
Topic1	european stock uk plc britain earnings top share market ftse
Topic2	canadian dollar yield bank rate canada india bond market cent
Topic3	opec market output cut production bpd producer supply future brent
Topic4	market dollar stock euro currency rate european yield bank bond
Topic5	billion bank company market fund saudi debt investor source investment
Topic6	stock share rose fell energy canada tsx pct main marker
Topic7	company billion share quarter production gas cost profit shale bp
Topic8	growth economy china inflation rate demand quarter increase forecast expect
Topic9	gas energy project company lng pipeline power coal industry market
Topic10	government saudi country president minister Qatar official arabia political iran
Topic11	stock rose nasdaq fell dow share market investor earnings fed
Topic12	refinery fuel gasoline supply harvey hurricane refiner demand coast import

表 2 pLSA 主题模型得到的相应主题下概率最高的十个词

主题	词
Topic0	stock billion company bank dollar gas market government production cut
Topic1	market stock bank rose rate rating fitch fell share company
Topic2	market rate company investor supply dollar canadian fell rose future
Topic3	company rate dollar fell day quarter cut share yield future
Topic4	bank company share billion energy stock investor gain analyst cut
Topic5	market stock billion energy company dollar rating investor share demand
Topic6	market stock share bank rating investor dollar rose day energy
Topic7	market stock bank company billion rate investor fund future analyst
Topic8	market cut stock dollar opec output saudi rate day bond
Topic9	rating market fitch bank billion dollar share low financial debt
Topic10	company market bank time cut country gas production output global
Topic11	bank dollar rose investor day market global stock time source
Topic12	market share investor dollar global stock energy canadian fitch time

与需求相关. 根据表 1~3 可以看出 LDA 和 HDP 模型划分的新闻主题, 由每个主题概率最高的前十个词基本能概括出该主题的相关内容.

采用上述三种模型对油价相关新闻进行主题识别, 可得到主题分布以及每一篇文档所对应概率最高的主题, 进而可以计算主题分类下的新闻情感. 计算过程大致为: 根据新闻所对应概率最高的主题将每一篇新闻划分到不同主题下, 并运用 Textblob 得到每一篇新闻的情感分值, 经过平均计算得到每一个主题下的平均新闻篇章情感, 13 个主题即为 13 个主题分布变量以及 13 个主题分类下新闻情感变量.

表 3 HDP 主题模型得到的相应主题下概率最高的十个词

主题	词
Topic0	market stock bank company dollar share rate billion energy rose
Topic1	rating fitch bank market security issuer risk government support debt
Topic2	market stock bank rate canadian canada rose fell dollar energy
Topic3	market stock bank rate dollar share yield fell currency company
Topic4	rating fitch opec market cut producer company output saudi energy
Topic5	market india bank company trump bond billion debt investor indian
Topic6	market cut stock bank opec day fell production company global
Topic7	dollar market rate fed nasdaq cut day opec majot stock
Topic8	production opec bpd shale energy market output rise cut minister
Topic9	market opec backwardation genel production airport gas dollar fund taq
Topic10	market cut dollar opec stock meeting investor fed currency china
Topic11	market demand cut copper production global output dollar gasoline metal
Topic12	Indonesia gas trump coal government production china country revenue bank

3.2.2 油价预测结果分析

根据上一小节得到的网络新闻主题分布和主题分类下的新闻情感分值, 分别加入到原油价格预测模型中, 来测试对原油价格预测模型的改进程度. 相关实证分析结果见表 4 至表 10. 这里采用水平预测精度指标 MAPE 和 RMSE 来评价不同模型在日度原油价格预测中的预测效果, 并对油价预测效果的改进程度 (用 IR 指标来测度) 进行比较分析.

首先, 运用不同预测模型对原油价格序列进行预测, 可得到原始油价的预测结果, 见表 4 所示. 通过比较可以看出: 1) LR 模型的预测结果最好, 而 RVFL 模型的预测效果最差. 主要原因可能在于时间跨度不够长的高频油价数据运用线性模型就能达到较好的短期预测效果; 2) 不同模型在同一评价标准下的表现有较大差别, 原因在于不同模型的算法不同, 在预测中的表现也不同.

表 4 基于不同模型的原油价格预测结果

模型	MAPE	RMSE
LR	0.0113	0.8278
RVFL	0.2282	17.1032
SVR	0.0299	1.9056
ELM	0.0222	1.4402

其次, 将 LDA 主题模型得到的主题分布加入油价预测模型中以测定对油价预测的影响, 预测效果的改进如表 5 所示.

从表 5 中可以看出: 1) LR 模型预测效果的改进最差. 主要原因可能是对于多元线性回归而言加入的变量越多, 变量间越可能存在共线性进而影响预测效果; 2) 除 LR 之外, 其它三个人工智能模型在加入主题分布后预测效果都得到了有效的提升, 即三个基于人工智能的非线性模型预测提升效果明显优于线性 LR 模型, 证实网络新闻主题分布包含的信息对预测效果提升的有效性, 且对非线性智能模型更加有效; 3) 在三个人工智能模型, RVFL 模型在加入 LDA 主题分布后预测效果改进最大, 其次是 SVR, 最后是 ELM. 也就是说, 主题分布对于不同人工智能模型的预测效果提升存在差异.

再次, 在新闻主题确定后, 可根据主题对新闻进行分类, 得到各个相应主题分类下的新闻

情感分值, 然后将其加入到预测模型以测定主题分类下新闻情感对油价预测结果的影响. 表 6 列出了加入 LDA 主题分类下新闻情感分值后的预测结果. 从表 6 不难看出: 1) 与加入新闻主题分布获得的结果类似, RVFL 的预测精度改进最大, LR 模型的改进结果最差. 可能原因同上; 2) 除 LR 外的其他三个人工智能模型的预测效果都得到了有效提升, 证实了主题分类下新闻情感信息对预测效果提升的有效性; 3) 与表 5 对比, 考虑主题分类下新闻情感比只考虑主题分布对油价预测效果的提升更大, 特别是在 SVR 模型与 ELM 模型的表现更加明显. 可能原因在于主题分类下网络新闻情感包含更多有价值的信息, 从而更能改进预测效果; 4) 无论是加入新闻主题分布还是加入主题分类下的新闻情感, 这些模型在同一评价标准下彼此之间的差异大大减少, 显示出了类似有效的预测能力, 表明加入对油价预测密切相关的信息, 就可能改进油价预测的效果, 证实了原油市场信息的有效性.

类似地, 将 pLSA 主题模型得到的主题分布加入不同油价预测模型中, 对预测效果的提升如表 7 所示. 同时, 加入对应主题分类下新闻情感分值后, 对预测效果的改进如表 8 所示.

根据表 7 与表 8 可以看出: 无论是加入新闻主题分布还是加入主题分类下的新闻情感, 除 LR 外的模型在预测效果上都有较大的改进, 且主题分类下的新闻情感在加入预测模型后有更大的预测效果提升. 在这两种情况下, 非线性模型较线性模型依旧表现出优势. 与 LDA 的预测效果对比可以看出, pLSA 模型得到的主题分布与主题分类下的情感分别加入油价预

表 5 加入 LDA 主题分布的预测效果

模型	MAPE	IR _{MAPE}	RMSE	IR _{RMSE}
LR	0.0142	-24.91%	1.0266	-24.01%
RVFL	0.0138	93.97%	1.0034	94.13%
SVR	0.0164	45.35%	1.1290	40.75%
ELM	0.0205	7.72%	1.3309	7.59%

表 6 加入 LDA 主题分类下情感的预测效果

模型	MAPE	IR _{MAPE}	RMSE	IR _{RMSE}
LR	0.0153	-35.04%	1.0776	-30.18%
RVFL	0.0138	93.96%	0.9994	94.16%
SVR	0.0152	49.23%	1.0732	43.68%
ELM	0.0186	16.57%	1.2200	15.29%

表 7 加入 pLSA 主题分布的预测效果

模型	MAPE	IR _{MAPE}	RMSE	IR _{RMSE}
LR	0.0130	-14.47%	0.9488	-14.61%
RVFL	0.0140	93.88%	1.0023	94.14%
SVR	0.0127	57.71%	0.9574	49.76%
ELM	0.0143	35.72%	0.9944	30.96%

表 8 加入 pLSA 主题分类下情感的预测效果

模型	MAPE	IR _{MAPE}	RMSE	IR _{RMSE}
LR	0.0128	-13.18%	0.9372	-13.22%
RVFL	0.0127	94.45%	0.9582	94.40%
SVR	0.0119	60.30%	0.8999	52.78%
ELM	0.0124	44.41%	0.8841	38.62%

测模型中, 预测效果的提升优于 LDA 模型. 可能的原因在于主题模型的构造不同, 主题划分不同进而导致在预测中的表现不同.

最后, 将 HDP 主题模型得到的主题分布加入油价预测模型中, 测试 HDP 主题模型得到的主题分布对油价预测效果改进的程度, 相关预测结果如表 9 所示. 同样, 加入对应主题分类下新闻情感分值的预测效果如表 10 所示.

从表 9 与表 10 可看出: 1) 无论是加入新闻主题分布还是加入主题分类下的新闻情感, SVR 模型都表现出较好的预测水平精度, 展示了 SVR 模型的全局优化能力; 2) 加入 HDP 新闻主题分布后, 除 LR 外的三个人工智能模型在预测效果上都有较大的改进, 非线性智能预测模型较线性 LR 预测模型依旧表现出较大优势; 3) 对比 HDP 新闻主题分布, 加入 HDP 主题分类下网络新闻情感后对油价预测效果的改进更大. 究其原因, 可能 HDP 主题分类下网络新闻情感包含更多有价值的信息, 从而改进预测效果; 4) 与 LDA、pLSA 不同的是, 加入 HDP 主题分类下的新闻情感预测后无论线性模型还是非线性模型预测效果都有改进. 主要原因可能在于主题模型的构造不同, 主题划分不同导致在预测中的表现不同.

为了更加直观地比较各模型之间的预测能力, 本文使用 DM 检验 (Diebold Mariano test) 对加入主题分布与主题分类下新闻情感后得到的预测结果与未加入新闻信息的预测结果进行对比, 具体结果如表 11 所示.

由表 11 的结果可以看出: 1) 在绝大部分情况下 3 种主题模型下 3 种非线性模型的检验结果显著, 表明将新闻主题分布与主题分类下的新闻情感在分别加入油价预测后能够较为有效的提升非线性模型的预测性能; 2) 对于线性模型 LR 来说, 加入更多的变量并不能显著提高其预测性能.

总的来说, 根据以上 3 种主题模型的结果对比可以得出, 对于非线性预测模型来说, 新闻主题分布与主题分类下的新闻情感在分别加入油价预测后都能够有效提升预测效果, 且主题分类下的新闻情感能更大程度上提升预测效果, 证明了主题分类下网络新闻情感对改进油价预测效果的有效性. 其主要原因在于主题划分的新闻篇章情感更强调每一个主题聚合下的

表 9 加入 HDP 主题分布的预测效果

模型	MAPE	IR _{MAPE}	RMSE	IR _{RMSE}
LR	0.0114	-0.75%	0.8396	-1.42%
RVFL	0.0152	93.34%	1.0390	93.93%
SVR	0.0105	64.92%	0.7798	59.08%
ELM	0.0181	18.49%	1.1954	17.00%

表 10 加入 HDP 主题分类下情感的预测效果

模型	MAPE	IR _{MAPE}	RMSE	IR _{RMSE}
LR	0.0112	0.81%	0.8244	0.41%
RVFL	0.0116	94.92%	0.8394	95.09%
SVR	0.0107	64.11%	0.7833	58.90%
ELM	0.0170	23.74%	1.1300	21.54%

表 11 加入不同模型主题分类或情感的 DM 检验结果

模型	加入 LDA 主题分类		加入 LDA 主题分类下情感	
	MAPE	MSE	MAPE	MSE
LR	0.2334	0.0997*	0.2548	0.3663
RVFL	0.0000***	0.0000***	0.0000***	0.0000***
SVR	0.0000***	0.0015***	0.0000***	0.0009***
ELM	0.5426	0.2607	0.0940*	0.0019***
	加入 pLSA 主题分类		加入 pLSA 主题分类下情感	
	MAPE	MSE	MAPE	MSE
LR	0.2334	0.0997*	0.2548	0.3663
RVFL	0.0000***	0.0000***	0.0000***	0.0000***
SVR	0.0000***	0.0015***	0.0000***	0.0009***
ELM	0.5426	0.2607	0.0940*	0.0019***
	加入 HDP 主题分类		加入 HDP 主题分类下情感	
	MAPE	MSE	MAPE	MSE
LR	0.2334	0.0997*	0.2548	0.3663
RVFL	0.0000***	0.0000***	0.0000***	0.0000***
SVR	0.0000***	0.0015***	0.0000***	0.0009***
ELM	0.5426	0.2607	0.0940*	0.0019***

注：*、**、*** 分别表示 10%、5% 和 1% 的显著性水平。其中原假设和备择假设分别为：H0：加入不同主题模型的主题分类或主题分类下情感后模型的预测能力与原模型无明显差异；H1：加入不同主题模型的主题分类或主题分类下情感后模型的预测能力优于原模型。

情感, 而每一个主题下的情感都是一个影响因素, 从而能够有效提升油价预测效果.

4 总结与展望

本文主要针对现有油价预测研究中存在的不足, 提出了融合网络新闻主题分布与主题分类下新闻情感的油价预测方法, 并进行了实验设计与实证研究. 具体而言, 通过主题识别模型和情感分析技术, 对包含大量有关事件和舆论因素的网络新闻进行主题识别, 将其转化为主题分布数据, 并通过主题对新闻情感进行分类, 之后将主题分布与每个主题下的新闻情感分别加入到原油价格预测模型中, 对预测效果的改善进行了比较分析. 实证结果表明: 与基准模型相比, 网络新闻主题分布与主题分类下的新闻情感在分别加入油价预测模型后都能够有效提升预测效果, 且主题分类下的新闻情感能更大程度上改进预测效果, 证实了网络新闻主题分布与主题分类下新闻情感对改进油价预测效果的有效性. 这也说明, 网络新闻主题以及

每个主题分类下的新闻情感可用于对油价预测的研究之中。

然而, 本文还存在如下可以进一步研究的问题, 如尝试采用更多的主题模型来划分主题并计算主题分布; 运用其他情感分析模型, 优化情感分析过程; 采用更多的线性与非线性预测模型来测试该方法对油价预测效果的影响; 尝试收集更多的网络新闻文本来进行实证。这些问题将在未来的研究中进一步开展。

参 考 文 献

- 陈海文, 蔡志平, 方峰, (2016). 应用财经新闻挖掘的金融品种价格走势预测 [J]. 计算机工程与科学, 38(9): 1909–1916.
- Chen H W, Cai Z P, Fang F, (2016). Financial Price Trend Forecast Using Financial News Mining[J]. Computer Engineering & Science, 38(9): 1909–1916.
- 陈晓红, 彭宛露, 田美玉, (2016). 基于投资者情绪的股票价格及成交量预测研究 [J]. 系统科学与数学, 36(12): 2294–2306.
- Chen X H, Peng W L, Tian M Y, (2016). Stock Market Prediction Based on Investor Sentiment[J]. Journal of Systems Science and Mathematical Sciences, 36(12): 2294–2306.
- 侯璐, (2009). 基于 ARIMA 模型的石油价格短期分析预测 [D]. 广州: 暨南大学.
- Hou L, (2009). The ARIMA Model Based on Short-term Analysis and Forecasting of Oil Prices[D]. Guangzhou: Jinan University.
- 赵沙, (2013). 统计模型在国际原油价格预测及天然气定价中的应用 [D]. 北京: 清华大学.
- Zhao S, (2013). Applying Statistical Models to International Crude Oil Price Forecasting and Natural Gas Pricing[D]. Beijing: Tsinghua University.
- 赵阳, (2015). 基于压缩感知的国际油价预测方法研究 [D]. 北京: 北京化工大学.
- Zhao Y, (2015). The Research of Compressed Sensing Based International Crude Oil Price Forecasting Approach[D]. Beijing: Beijing University of Chemical Technology.
- 郑俊艳, (2012). 国际原油价格预测模型研究 [D]. 天津: 天津大学.
- Zheng J Y, (2012). The Research of International Crude Oil Price Forecasting Model[D]. Tianjin: Tianjin University.
- Aharon D Y, Demir E, Lau C K M, Zaremba A, (2022). Twitter-Based Uncertainty and Cryptocurrency Returns[J]. Research in International Business and Finance, 59: 101546.
- Alfano S J, Feuerriegel S, Neumann D, (2015). Is News Sentiment More than Just Noise?[C]// Twenty-Third European Conference on Information Systems (ECIS).
- Baumeister C, Kilian L, (2012). Real-time Forecasts of the Real Price of Oil[J]. Journal of Business & Economic Statistics, 30(2): 326–336.
- Blei D M, Ng A Y, Jordan M I, (2003). Latent Dirichlet Allocation[J]. Journal of Machine Learning Research, 3: 993–1022.
- Brey T, Jarre-Teichmann A, Borlich O, (1994). Artificial Neural Network Versus Multiple Linear Regression: Predicting P/B Ratios from Empirical Data[J]. Marine Ecology Progress Series, 140: 251–256.
- Burges C J C, (1998). A Tutorial on Support Vector Machines for Pattern Recognition[J]. Data Mining and Knowledge Discovery, 2(2): 121–167.
- Coppla A, (2008). Forecasting Oil Price Movements: Exploiting the Information in the Futures Market[J]. Journal of Futures Markets, 28(1): 34–56.

- Cortes C, Vapnik V, (1995). Support-vector Networks[J]. *Machine Learning*, 20(3): 273–297.
- Griffiths T L, Steyvers M, (2004). Finding Scientific Topics[J]. *Proceedings of the National Academy of Sciences*, 101(Suppl_1): 5228–5235.
- Heinrich G, (2005). *Parameter Estimation for Text Analysis*[R]. Leipzig: University of Leipzig.
- Hofmann T, (1999). Probabilistic Latent Semantic Analysis[C]// *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann Publishers Inc: 289–296.
- Huang G B, Zhu Q Y, Siew C K, (2006). *Extreme Learning Machine: Theory and Applications*[J]. *Neurocomputing*, 70(1): 489–501.
- Ji Q, Guo J F, (2015). Oil Price Volatility and Oil-related Events: An Internet Concern Study Perspective[J]. *Applied Energy*, 137: 256–264.
- Kaboudan M A, (2001). Compumetric Forecasting of Crude Oil Prices[C]// *Proceedings of the 2001 Congress on Evolutionary Computation (IEEE Cat. No. 01TH8546)*. IEEE, 1: 283–287.
- Kraaijeveld O, De Smedt J, (2020). The Predictive Power of Public Twitter Sentiment for Forecasting Cryptocurrency Prices[J]. *Journal of International Financial Markets, Institutions and Money*, 65: 101188.
- Kullback S, Leibler R A, (1951). On Information and Sufficiency[J]. *The Annals of Mathematical Statistics*, 22(1): 79–86.
- Lechthaler F, Leinert L, (2012). Moody Oil: What is Driving the Crude Oil Price?[J]. *Empirical Economics*: 1–32.
- Li J, Xu Z J, Yu L, Tang L, (2016). Forecasting Oil Price Trends with Sentiment of Online News Articles[J]. *Procedia Computer Science*, 91: 1081–1087.
- Li X, Shang W, Wang S, (2017). Crude Oil Price Movement Volatility Forecasting Based on Online News[C]// *21th Pacific Asia Conference on Information Systems (PACIS)*, 164.
- Lu X, Ma F, Xu J, Zhang Z, (2022). Oil Futures Volatility Predictability: New Evidence Based on Machine Learning Models1[J]. *International Review of Financial Analysis*, 83, 102299.
- Mirmirani S, Li H C, (2004). A Comparison of Var and Neural Networks with Genetic Algorithm in Forecasting Price of Oil[J]. *Advances in Econometrics*, 19(4): 203–223.
- Morana C, (2001). A Semiparametric Approach to Short-term Oil Price Forecasting[J]. *Energy Economics*, 23(3): 325–338.
- Murat A, Tokat E, (2009). Forecasting Oil Price Movements with Crack Spread Futures[J]. *Energy Economics*, 31(1): 85–80.
- Naseer A, Nava Baro E, Daud Khan S, Vila Y, Doyle J, (2022). A Novel Cryptocurrency Prediction Method Using Optimum CNN[J]. *Computers, Materials & Continua*, 71(1): 1051–1063.
- Pao Y H, Park G H, Sobajic D J, (1994). Learning and Generalization Characteristics of the Random Vector Functional-link Net[J]. *Neurocomputing*, 6(2): 163–180.
- Skuza M, Romanowski A, (2015). Sentiment Analysis of Twitter Data Within Big Data Distributed Environment for Stock Prediction[C]// *2015 Federated Conference on Computer Science and Information Systems (FedCSIS)*. IEEE: 1349–1354.
- Smales L A, (2016). Time-varying Relationship of News Sentiment, Implied Volatility and Stock Returns[J]. *Applied Economics*, 48(51): 4942–4960.
- Tang L, Wu Y, Yu L, (2018). A Non-iterative Decomposition-ensemble Learning Paradigm Using RVFL Network for Crude Oil Price Forecasting[J]. *Applied Soft Computing*, 70: 1097–1108.

- Teh Y W, Jordan M I, Beal M J, Blei D M, (2006). Hierarchical Dirichlet Processes[J]. Journal of the American Statistical Association, 101(476): 1566–1581.
- Urolagin S, Sharma N, Datta T K, (2021). A Combined Architecture of Multivariate LSTM with Mahalanobis and Z-Score Transformations for Oil Price Forecasting[J]. Energy, 231: 120963.
- Wex F, Widder N, Liebmann M, Neumann D, (2013). Early Warning of Impending Oil Crises Using the Predictive Power of Online News Stories[C]// 2013 46th Hawaii International Conference on System Sciences. IEEE: 1512–1521.
- Yu L, Dai W, Tang L, (2015). A Novel Decomposition Ensemble Model with Extended Extreme Learning Machine for Crude Oil Price Forecasting[J]. Engineering Applications of Artificial Intelligence, 47: 110–121.
- Zhang L, Suganthan P N, (2016). A Comprehensive Evaluation of Random Vector Functional Link Networks[J]. Information Sciences, 367: 1094–1105.
- Zhang X, Yu L, Wang S, Lai K, (2009). Estimating the Impact of Extreme Events on Crude Oil Price: An EMD-based Event Analysis Method[J]. Energy Economics, 31(5): 768–778.