

网络舆情群体观点提取模型构建与实证研究

——以生猪期现货网络舆情为例

赵江元¹ 王伟军¹ 黄 微²

(1. 大连商品交易所, 辽宁 大连 116000; 2. 吉林大学商学与管理学院, 吉林 长春 130012)

摘要: [目的/意义] 本文提出并构建群体观点提取模型, 划分网络舆情群体观点的同时提取重点关键词展现群体观点的实质内涵, 为舆情引导提供参考。[方法/过程] 基于 LDA 主题模型与 TextRank 关键词模型构建群体观点提取模型, 以生猪+期货为关键词抓取网络舆情数据为实验样本, 计算最佳群体观点数量后进行群体观点划分, 并提取核心关键词来体现观点的语义内涵。[结果/结论] 基于 LDA 和 TextRank 的群体观点提取模型, 能够较好地根据语义内涵划分群体观点并提取核心关键词, 经过提取并排序后的关键词能够明确体现各个群体观点的核心语义内涵, 有助于服务金融和舆情监管机构根据期现货市场行情、群体观点语义内涵以及数量峰值出现阶段合理开展舆情监控与引导工作。

关键词: 网络舆情; 期货市场; 群体观点; 关键词提取

DOI: 10.3969/j.issn.1008-0821.2024.02.013

[中图分类号] G203 [文献标识码] A [文章编号] 1008-0821 (2024) 02-0152-08

Construction and Empirical Study of a Group Opinion Extraction Model for Online Public Opinion

——A Case Study of Live Pig Spot Online Public Opinion

Zhao Jiangyuan¹ Wang Weijun¹ Huang Wei²

(1. Dalian Commodity Exchange, Dalian 116000, China;

2. School of Business and Management, Jilin University, Changchun 130012, China)

Abstract: [Purpose/Significance] This paper proposes and constructs a group view point extraction model, which divides the opinions of online public opinion groups while extracting key keywords to demonstrate the essence of group viewpoints, providing reference for public opinion guidance. [Method/Process] Based on the LDA topic model and TextRank keyword model, this paper constructed a group opinion extraction model and captured the online public opinion data with pig+futures as the key words as the experimental sample. After calculating the optimal number of group opinions, the study divided group opinions and extracted core keywords to reflect the semantic connotation of opinions. [Result/Conclusion] The group viewpoint extraction model based on LDA and TextRank can effectively classify group viewpoints based on semantic connotations and extract core keywords. After extraction and sorting, the keywords can clearly reflect the core semantic connotations of each group viewpoint, which is helpful for serving financial and public opinion regulatory agencies to carry out reasonable public opinion monitoring and guidance work based on the spot market situation, semantic connotations of group viewpoints, and the stage of peak quantity.

Key words: online public opinion; futures market; group perspective; keyword extraction

收稿日期: 2023-05-11

作者简介: 赵江元 (1993-), 男, 博士后, 研究方向: 网络舆情。王伟军 (1969-), 男, 副总经理, 研究方向: 资本市场、新闻信息传播。黄微 (1967-), 男, 博士生导师, 研究方向: 网络舆情。

随着网络时代的发展,网络舆情已经成为社会舆论的一种新的表现形式,众多网民的参与使网络舆情呈现从众化、去个性化、群体极化等特点,网络舆情的演化与传播极易受到群体观点的影响。同时,随着互联网规模的持续增长,网络舆情的生产速度、内容形式、传播态势等也发生重大变化,网络环境愈发繁杂,网络舆情群体观点的内隐性逐步显现,通过常规手段已经较难从海量网络舆情信息中对网络舆情群体观点进行有效辨别,这对网络舆情监管机构实时、快速地响应复杂舆情危机带来了巨大挑战。因此,快速洞悉网络舆情群体观点的核心内涵是网络舆情监管机构实现舆情快速处置的重要手段。

本文构建基于 LDA+TextRank 的网络舆情群体观点提取模型,从实践层面对网络群体观点的分类与挖掘进行研究,并以期货市场中生猪期货相关的期现货市场的网络舆情信息作为研究样本,依靠困惑度指标确定每个时间段内网络舆情的最佳群体观点数量,划分群体观点并按照语义重要性对关键词进行重新排序,利用核心关键词表示群体观点的主要内涵。相关研究结果能够作为金融监管机构合理引导舆情、靶向监管的重要参考依据,为营造期货市场良好外部环境提供支持。

1 相关研究

1.1 网络舆情群体观点挖掘相关研究

观点挖掘(Opinion Mining)是指通过对用户发表评论的文本内容进行分析,提炼出用户对产品、服务等内容的情感、态度、观点等^[1]。目前较为普遍是将情感作为网络舆情观点的表达,如吴思等^[2]利用 Bert、Word2vec 等模型对当前较为流行的短视频类多维度舆情信息分别进行文本、声音、图像等维度数据提取,并利用卷积神经网络的 Softmax 层对多模态信息进行情感分类。王萍等^[3]提出一种包含上下文位置潜在信息的方面级情感分析模型,利用上下文位置调整函数设置上下文词的位置权重来增强情感分析的结果,最终在医疗文本上取得了较好的效果。随着研究的逐渐深入,网民对社会热点事件的评价不仅仅是针对事件本身表达出的褒义和贬义情感来体现,更多的是以某种具体的“看法”来体现其态度。因此,针对网络舆情、社会

评论等进行观点分类及语义内涵挖掘的研究逐步涌现。睢国钦等^[4]提出一种基于双向长短期记忆神经网络和 CRFS(条件随机场)的观点抽取方法,对电商平台评论集进行观点挖掘,取得了较好的效果。Fresneda J E 等^[5]采用潜在语义分析(Latent Semantic Analysis, LSA)研究了评论词汇,从语义层面研究了观点熵、确证熵、推荐熵对评论舆情有用性的影响。卢国强等^[6]基于 Word2vec 和 K-means 模型构建极端舆情观点识别模型,通过分析极端观点特征来对群体极化风险下的极端舆情观点进行识别。

1.2 网络舆情对资本市场影响的研究

信息是资本市场构成要素之一,网络舆情信息的传播对资本市场的运行情况及效率具有显著影响。Yang C 等^[7]研究发现,网络评论的情绪对股票期权具有一定预测作用,看涨情绪将对期权价格的市场预期产生正向影响,反之则产生负向影响。Smales L A^[8]通过分析 2003—2012 年的新闻信息情绪与黄金期货回报率之间关系发现,负面新闻情绪较多时黄金期货回报率更大。Gong X 等^[9]研究发现,在原油期货波动预测模型中加入通过从网络信息中提取的投资者情绪指标,预测能力显著提升,模型拟合效果得到有效改善。Lin C B 等^[10]通过对投资者情绪与期现货市场定价能力之间的关系进行了研究,发现从网络信息中收集的多空观点能够对期现基差波动产生一定的正面影响,在看多观点较多时则期货市场的短期定价主导作用减弱。Seo S W 等^[11]提出一种基于期权隐含信息的波动率预测模型,并通过实证发现,在网络舆情情绪高涨阶段模型对期权交易回报的可预测性较低迷情绪期间要更强。

综合来看,通过从语义层面对网络舆情的群体观点进行提取是目前网络舆情研究的主要方向。同时,网络舆情对资本市场的波动性、收益率也确实会造成一定影响。目前研究多是从投资者情绪即情感维度来探讨多空情绪或积极、消极情绪网络舆情对资本市场的影响。从资本市场监管角度来看,缓解舆情影响的关键在于合理引导舆论走向,单纯对网络舆情情感的分析而不能洞悉网络舆情中核心观点难以精准掌握舆情引导的切入点,同时结合目前网络舆情研究趋势,对网络舆情进行观点提取是未来网络舆情领域研究的一个重点。因此,本文通过

结合 LDA 和 TextRank 构建的群体观点提取模型,对期货市场网络舆情信息中的群体观点及其语义内涵进行挖掘,是为资本市场舆情引导提供切入点的有效方式。

2 群体观点提取模型构建

2.1 群体观点提取模型构建过程

群体观点提取模型的核心功能在于,从观点内涵的层面快速精准对某个时间段内网络舆情信息进行分类,提取出关键词的同时根据关键词对群体观点内涵的表示效力来进行重新排序,反映当前网络舆情中各类群体观点所表达的主要思想,从而实现网络舆情群体观点的快速理解。

LDA 模型虽然也能对关键词进行提取,但该模型是通过计算词的分布和文档分布的相似度,取词属于主题的概率作为关键词选取的依据,更侧重于强调词的主题属性,而忽略了词上下文相关词对

该词在主题中的语义体现。TextRank 是一种基于图形的文本处理排序算法,可以更充分地利用文档集中各词之间的关系,且经过排序后的关键词集合可以作为快速识别网络舆情群体观点的语义内涵。因此,本文基于 LDA 的主题分类功能与 TextRank 的关键词提取排序功能相结合来构建群体观点提取模型。具体构建过程如下:利用网络爬虫抓取文本语料,以每天为时间窗口将当日网络舆情文本汇聚为一个文档集,通过计算困惑度得出文档的最佳群体观点分类数量,输入至利用 LDA 模型对群体观点进行划分,随后通过 TextRank 对分属各个群体观点的网络舆情数据集进行关键词提取并排序,最终依靠排序后的前几位关键词即可作为各个网络舆情群体观点的语义描述。群体观点提取模型构建过程如图 1 所示。

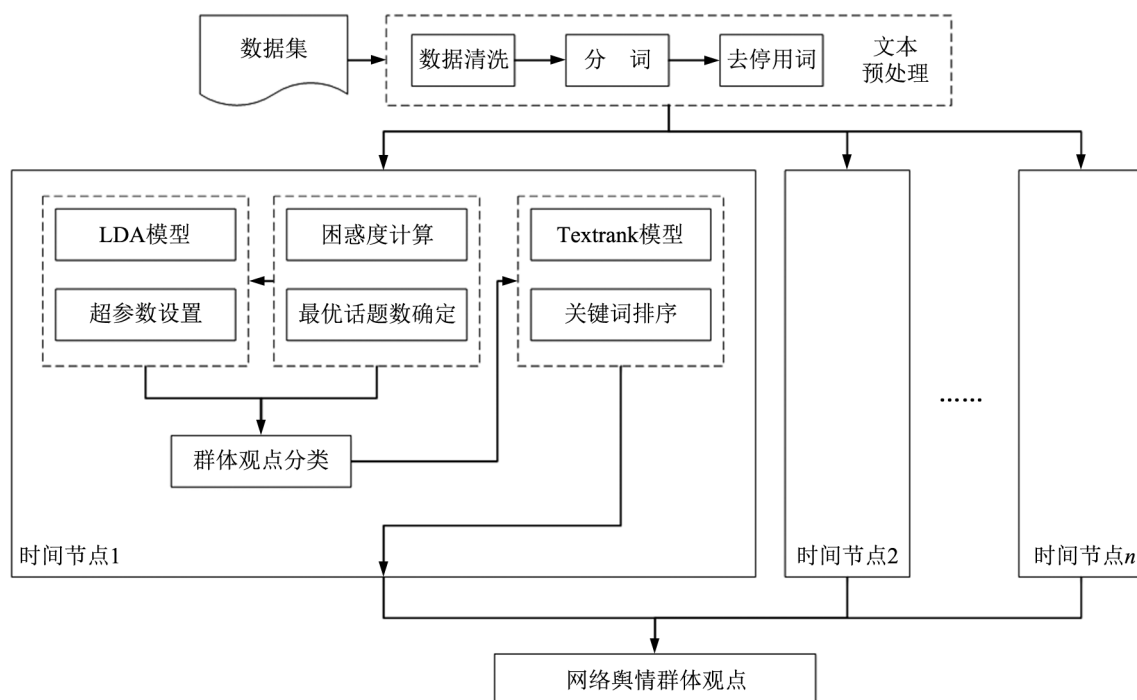


图 1 网络舆情群体观点提取流程图

Fig. 1 Flowchart for Extracting Group Opinions in Online Public Opinion

2.2 网络舆情群体观点提取方法

LDA(隐狄利克雷分配模型, Latent Dirichlet Allocation)^[12]作为一种无监督的概率主题模型,能够较好地将文档主题以概率的形式表现出来,可以用来识别大规模文档集或语料库中的潜藏主题信息。LDA 是目前较为主流的主题模型,在对网络舆情群体观点的分类上具有较好的效果。同时,将主题词

与分类情况相结合,可以较好地反映网络舆情中群体观点的深层次语义内涵。

LDA 的基本原理是^[12]:假定形成群体观点的信息共 M 条,第 i 条舆情信息包含 n_i 个关键词,模型输出为每条舆情信息的群体观点分类与每个群体观点中关键词的分布,假定每个时间段内舆情语料中共包含 K 个话题。由于在贝叶斯推断中, Dirichlet

分布作为多项分布的共轭先验应用，在 LDA 模型中假定网络舆情群体观点先验分布是 Dirichlet 分布，即对任意网络舆情语料 i 其群体观点分布为 θ_i ，则 θ_i 如式 (1) 所示：

$$\theta_i = \text{Dirichlet}(\vec{\alpha}) \quad (1)$$

式 (1) 中， α 为主题分布的超参数，是一个 K 维向量，用以控制网络舆情语料中 K 个群体观点，根据以往经验， α 一般设置为 $100/K$ 。针对微博 i 中的每个词 j ，可以从群体观点分布 θ_i 中采样生成关键词 j 的观点分类。

同样，假定网络舆情群体观点中关键词的先验分布是 Dirichlet 分布，则群体观点—关键词分布 $\phi_{z_i,j}$ 如式 (2) 所示：

$$\phi_{z_i,j} = \text{Dirichet}(\vec{\beta}) \quad (2)$$

式 (2) 中， β 为关键词分布的超参数，是一个 z 维向量， z 为所有网络舆情语料中关键词的数量， β 一般设置为 0.01。从群体观点—关键词分布 $\phi_{z_i,j}$ 中采样生成关键词 $w_{i,j}$ 。

最终，一条网络舆情信息中关键词分布 $p(w_i | \alpha, \beta)$ 的最大似然值估计可以通过将 θ_i 以及 ϕ 进行积分和对 z_i 进行求和得到：

$$p(w_i | \alpha, \beta) = \int \int \sum_{z_i} p(w_i, z_i, \theta_i, \phi | \alpha, \beta) \quad (3)$$

式 (3) 中 $p(w_i | \alpha, \beta)$ 的最大似然估计，可以通过 Gibbs 采样方法估计出模型参数。Gibbs 采样算法初始化是随机将每个词分配给一个群体观点，通过多次迭代后，Gibbs 采样过程收敛，即得到最终的群体观点概率分布，迭代完成后输出网络舆情语料—群体观点矩阵和群体观点—关键词概率矩阵。根据每条网络舆情信息属于群体观点概率大小即可构建各个群体观点的网络舆情信息集合，该集合将作为 TextRank 的输入数据集用于提取群体观点的语义关键词。

2.3 最优群体观点数量确定

最优群体观点数量是影响网络舆情群体观点提取效果的关键因素，LDA 模型进行群体观点提取过程中最重要的问题是语料库中群体观点数量 K 值的确定，因此需要通过引入困惑度 (Perplexity)^[13] 指标来确定群体观点的最优个数。困惑度用于度量概率分布或概率模型预测样本的优劣程度，通过对比两

个概率分布或概率模型在预测样本上的优劣来选取最优模型。困惑度在评价聚类分类算法的性能上有极好的效果，因此可以通过对比困惑度来选取 LDA 的最优话题数。困惑度针对不同模型从概率分布困惑度、概率模型困惑度以及分词困惑度 3 种方法计算困惑度，针对自然语言处理模型通常选用 Perplexity Per Word，即分词困惑度方法进行困惑度计算。在语料集 D_i 上，困惑度表达式如式 (4) 所示：

$$\text{Perplexity}(D_i) = \exp \left\{ \frac{\sum_{d=1}^M \log p(w_d | M)}{\sum_{d=1}^M N_d} \right\} \quad (4)$$

其中， M 是指已训练模型参数，在 LDA 模型中为网络舆情语料—群体观点矩阵和群体观点—关键词概率矩阵， N_d 为网络舆情信息 d 中关键词数量， w_d 为语料集 D_i 中网络舆情信息 d 的词向量形式。由公式可知，中括号内分子项为测试集中网络舆情信息 d 属于模型上的概率，其值越大代表困惑度越小，说明模型的分类性能越高。通过对比选取不同群体观点数量 K 的 LDA 模型的困惑度作为最优群体观点数量判断依据，为避免选取的 K 值过大导致划分的群体观点语义内涵不聚焦，因此选取困惑度变化曲线第一次发生较大变化的拐点数值作为最优群体观点数量。

在实际操作时需要自动识别最佳群体观点数量，但离散点组成的曲线无法直接提取最佳群体观点数量。因此，本文采用曲线拟合的方式，将离散点组成的函数拟合成为连续函数并求其二阶导数，利用函数曲线二阶导数为零则为函数拐点的性质，将困惑度拟合曲线第一个最低点作为网络舆情群体观点最优数量的值。困惑度拟合曲线示例如图 2 所示。

2.4 关键词提取与排序

TextRank 算法是一种用于关键词和文档摘要提取的排序算法，由 PageRank 算法改进而来，通过利用一篇文档内各词汇间共现关系提取关键词^[14]。TextRank 的主要思路为^[15]：首先将文本数据中的词语或句子进行分词，将词作为节点，计算节点间的相似度确定节点构成的边，并将所有节点间相似度数作为权值构成图模型网络。以矩阵的形式表示文本向量化的结果，并对该矩阵进行迭代计算，

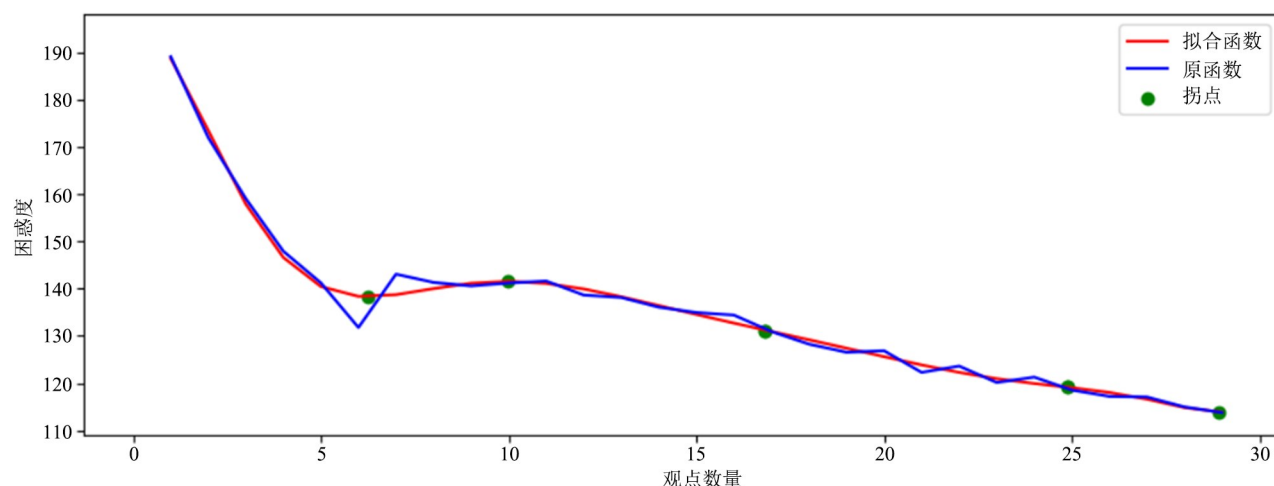


图 2 困惑度拟合曲线示例图
Fig. 2 Example Graph of Confusion Fit Curve

求得节点的权值大小,并根据节点的权值通过排序确定能够体现文本内容主题的词汇作为文档的关键词集合。TextRank 可以表示为一个有向有权图 $G=(V,E)$,由点集合 V 和边集合 E 组成, E 是 $V \times V$ 的子集。图中任意两点 V_i, V_j 之间边的权重为 w_{ji} ,对于一个给定的点 V_i ,点 V_i 的得分定义如式(5)所示:

$$WS(V_i) = (1-d) + d \times \sum_{V_j \in in(V_i)} \frac{w_{ji}}{\sum_{V_k \in out(V_i)} w_{jk}} WS(V_j) \quad (5)$$

其中, d 为阻尼系数,取值范围为 $0 \sim 1$,代表从图中某一特定点指向其他任意点的概率,一般取值为 0.85 ; $in(V_i)$ 为指向该点的点集合, $out(V_i)$ 为点 V_i 指向的点集合。使用 TextRank 算法计算图中各点的得分时,需要给图中的点指定任意的初值,并递归计算直到收敛,即图中任意一点的误差率小于给定的极限值时就可以达到收敛,一般该极限值取 0.0001 。

基于 TextRank 的关键词提取就是从一段给定的文本中自动抽取出若干有意义的词语或词组。TextRank 算法是利用局部词汇之间关系(共现窗口)对后续关键词进行排序,直接从文本本身抽取。其主要步骤如下:把给定的文本 T 按照完整句子进行分割,即 $T=[S_1, S_2, S_3, \dots, S_m]$ 。对于每个句子 $S_i \in T$,进行分词和词性标注处理并过滤掉停用词,只保留指定词性的单词,如名词、动词、形容词,即 $S_i=[V_{i1}, V_{i2}, V_{i3}, \dots, V_{in}]$ 。然后构建候选关键词图 $G=(V,E)$,其中 V 为节点集,由之前生成的候

选关键词组成,然后采用共现关系构造任意两点之间的边,仅当它们对应的词汇在长度为 N 的窗口中共现, N 表示窗口大小,即最多共现 N 个单词。根据式(5),迭代传播各节点的权重,直至收敛。对节点权重进行倒序排序,从而得到最重要的 T 个单词,作为候选关键词。将得到最重要的 T 个单词,在原始文本中进行标记,若形成相邻词组,则组合输出为多词关键词。

3 实证研究

3.1 数据源选择及数据采集

为验证本文群体观点模型在网络舆情语义内容划分与识别上的适用性,帮助金融监管机构快速响应网络舆情,本文以 2022 年 1 月 1 日—12 月 31 日生猪期货网络舆情数据为实验样本,利用网络爬虫抓取微博、东方财富股吧中的评论作为数据来源,通过实证方式来验证群体观点提取模型对于复杂网络舆情文本的观点提取能力,为金融监管机构开展舆情介入与引导提供支持。2022 年 10 月开始,国内生猪受疫情反复、供需错配、出栏量增多等因素影响价格持续走低,下跌周期延续至年末,生猪期货主力合约价格从最高点的 22 500 元/吨跌至 2022 年年末的 17 200 元/吨,跌幅达到 23.6%,这种跌幅对于生猪这类事关国际民生的大宗商品影响巨大,而价格的持续下跌也使网络上产生了部分不利于生猪期货功能发挥的新闻信息,对国家宏观调控政策的实施以及舆情监管机构开展舆情引导产生了一定影响。本文以天为单位,利用 Gooseeker 爬虫

工具对 2022 年全年的生猪相关的网络舆情数据进行抓取,考虑到生猪期货的价格发现功能是产生于现货市场的交易,因此本文将网络舆情采集关键词设置为“生猪”+“生猪期货”,从微博、东方财富

网生猪吧共抓取网络舆情数据 31 349 条,主要字段包括日期、时间、博文、发布地址等。生猪期货主力合约 2022 年价格走势如图 3 所示。



图 3 生猪期货主力合约 2022 年价格走势

Fig. 3 Price Trend Chart of Main Contract for Live Pig Futures in 2022

3.2 数据处理与分析

在数据处理阶段,以 Pycharm 作为开发环境,Python 版本 3.7。首先对抓取网络舆情数据进行预处理,包括去除“转发微博”“网页链接”以及网址、表情乱码等无用字段,并以日期为时间节点切分数据集并利用 Jieba 分词对语料进行分词、去停用词后存入数据库中,作为后续处理的输入数据。

在分词过程中需要将如生猪期货、主力合约、基差等部分特有名词以及特定词添加到自定义词典中以保证分词的准确性。在进行话题提取前需确定各时间点的最优话题数,采用前文的困惑度计算方法得出各时间点最优群体观点数量。图 4 为生猪期货网络舆情群体观点数量和 7 日内价格涨跌幅的对比图。

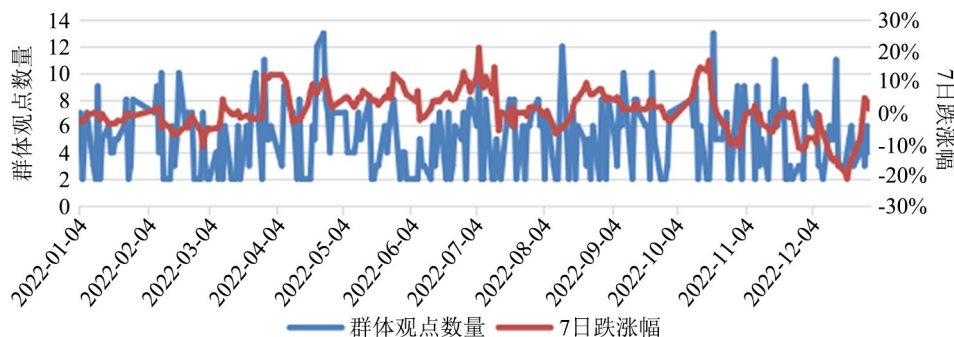


图 4 群体观点数量与 7 日内价格涨跌幅对比图

Fig. 4 Comparison Chart of Group Opinion Quantity and Price Fluctuations Within 7 Days

金融市场价格涨跌幅与网络舆情之间是相互影响的关系^[16],一方面,网络上的市场信息越多,越容易使价格产生变动;另一方面,价格大幅变化同样会激发更多网络舆情信息。如图 4 所示,生猪期货主力合约价格涨跌趋势与群体观点数量可以看出,如果某个交易日的 7 日价格涨跌幅度变化较大,则该交易日前后 3 个交易日内通常会出现一波群体观点数量高峰,即价格上涨引起网络中关于期现货市场价格的讨论明显增加,或市场突发新闻激发网络讨论的同时也带动价格大幅变动,如表 1 所示。

表 1 生猪期货价格涨跌幅与群体观点数量

Tab. 1 Price Fluctuations of Live Pig Futures and Quantity of Group Opinions

交易日	7 日价格 涨跌幅 (%)	群体观点 数量峰值 交易日	3 个交易日 内群体观点 数量峰值
3 月 29 日	12.11	3 月 29 日	11
4 月 25 日	10.62	4 月 25 日	13
7 月 5 日	21.01	7 月 5 日	11
12 月 20 日	-21.35	12 月 15 日	11

根据最佳观点数量利用 LDA 对每日网络舆情语料进行群体观点划分,然后通过 TextRank 对划分好的子观点进行关键词提取,TextRank 利用词权重构建图模型的思路对语料中词汇进行重新排序来体现词的重要性,因此结合关键词可以推断出各个群体观点的语义内涵,本文选取表 1 中出现群体观点峰值的各交易日权重前三的网络舆情群体观点,结合其部分关键词推断其观点语义内涵,为舆情引导提供指引和参考。部分交易日前三群体观点关键词及观点内涵如表 2 所示。

表 2 部分交易日前三群体观点关键词及观点内涵表
Tab. 2 Key Keywords and Content of Opinions from the Top Three Group Opinions on Selected Trading Days

日 期	部分关键词	群体观点语义内涵
3 月 29 日	双汇发展 2021 年 同比 生猪 公司 猪价 净利润 减少 9.72% 下降 收入……	上市公司“双汇发展”发布 2021 年年报
	农业 工作 疫情 防控 保障 菜篮子 稳定 生产 蔬菜 生猪 市场 落实……	各地保障疫情防控期间肉类蔬菜供应
	猪肉 价格 生猪 18.36 元 公斤 猪价 全国 下跌 监测 内蒙古 全省……	全国生猪价格下跌范围全线扩大
4 月 25 日	物价 中国 全球 近日 发改委 专家 座谈会 生猪 运行 平稳……	发改委召开发布会就生猪价格进行表态
	市场 价格 猪肉 生猪 24 元 每公斤 赔钱 养殖户 上涨 空间 有限……	市场猪肉价格上涨空间有限
	合肥 菜篮子 生猪 充足 供应 蔬菜 价格 猪肉 保供 稳产 同比……	合肥全力保障市民菜篮子稳定
7 月 5 日	生猪 发改委 价格司 猪肉 稳价 大连商品交易所 期货市场 保供 做好 研究	发改委表态做好期现联动监管
	上涨 国家 储备 生猪 近期 过快 发改委 猪肉 价格 投放 防范	发改委发文投放生猪储备抑制价格上涨
	发改委 生猪 价格 市场 上涨 持续 近期 行业 协会 会议 稳价 适时 储备	发改委召开会议讨论生猪价格上涨原因
12 月 15 日	国家统计局 11 月份 国民经济 处于 稳定 区间 猪肉 价格 不具有 上涨 基础	统计局表态生猪价格不具有上涨基础
	12 月 监测 显示 市场 生猪 价格 下降 猪肉 出栏 下跌	市场机构分析生猪价格趋势
	养殖 生猪 公司 建设 12 月 产能 价格 用电 排查 检查 供电	养殖相关信息汇总

从表 1 中可以看出,交易日 3 月 29 日、4 月 25 日、7 月 5 日的当日,群体观点数量都达到了阶段性峰值,以上 3 个交易日中生猪期货价格较 7 日前价格都出现了明显的上涨。因此交易日当日的多数群体观点的语义内涵都在强调生猪价格上涨空间有限、宏观管理部门表态要积极保供稳价等,属于价格变动产生网络信息。在这种情况下,金融监管机构则可以根据主要群体观点的语义内涵,开展具有针对性、方向性的引导工作,进一步降低市场预期、缓解生猪价格上涨趋势。

交易日 12 月 15 日权重最大的群体观点则是国家统计局表态生猪价格不具有上涨基础,在随后的 12 月 20 日,生猪期货价格较前 7 个交易日下跌-21%,这种情况则是国家部门发布权威市场信息导致价格下跌,由网络信息引发价格变动,金融监管部门则可以在重要信息在网络传播时提前开展舆情引导,强化国家部门发布权威信息对市场的引导和调节作用。

3.3 实验评估

主题模型作为一种无监督数据分析方法，通过对主题划分结果的评估，可以进一步对模型进行优化，使主题划分效果得到提升。因此对主题效果划分效果评估的评估十分重要。主题凝聚性度量方法 (Topic Coherence Measure) 是当前较为主流的主题聚类评估方法^[17-18]，用于评估每个模型生成的话题，其计算方式较为简单，适用于绝大多数主题模型效果评估^[19]。主题凝聚性指标定义如式 (6) 所示：

$$C(t; V^{(t)}) = \sum_{m=2}^M \sum_{l=1}^{m-1} \log \frac{CODF(v_m^{(t)}, v_l^{(t)}) + 1}{DF(v_l^{(t)})} \quad (6)$$

其中， $V^{(t)} = (v_1^{(t)}, \dots, v_m^{(t)})$ 是最能代表子话题 t 语义的 M 个关键词， DF 代表包含词 $v_l^{(t)}$ 的文档个数， $CODF$ 代表包含词 $v_l^{(t)}$ 和 $v_m^{(t)}$ 的文档个数。主题凝聚性系数 C 的值越高，则说明主题模型生成的话题质量越高，关键词对话题具有更好的解释效力。

因此，本文利用主题凝聚性度量方法来对各个模型抽取的关键词代表性进行评估，除 LDA、LDA+TextRank 以外，概率潜在语义分析模型 (PLSA, Probabilistic Latent Semantic Analysis) 作为 LDA 模型的前身，加入 PLSA 与本文模型一起进行主题凝聚性评估来体现本文模型的效果。LDA 与 PLSA 是给出关键词属于主题的概率来体现词的代表性，LDA+TextRank 是在 LDA 划分群体观点的基础上，利用词权重对词进行重新排序来体现词的代表性，因此 3 个模型都能给出每个群体观点中各个词的重要性，但作为舆情监控的参考要素，关键词过多将扰乱词对每个群体观点的代表性，因此选取各个模型提取的所有关键词中前 50 个关键词作为最能代表子话题语义的关键词合集进行主题凝聚性测度。各模型提取关键词的主题凝聚性如表 3 所示。

表 3 各模型主题凝聚性结果表
Tab. 3 Topic Coherence Results for Each Model

模 型	主题凝聚性
PLSA	-174
LDA	-167
LDA+TextRank	-159

通过表中结果可以看出，根据前文解释，主题凝聚性指数越大则主题凝聚性越高，本文的 LDA+

TextRank 的主题凝聚性指数数值要大于 PLSA 和 LDA，表明本文模型提取的关键词对生猪期货网络舆情群体观点的代表性要更加优越，关键词词组对群体观点的解释效力也更强。

4 结 语

本文基于 LDA 模型与 TextRank 模型构建网络舆情群体观点提取模型，以“生猪+期货”为关键词，抓取 2022 年相关网络舆情数据进行群体观点提取，并选取生猪期货主力合约价格涨跌幅度较大的交易日内网络舆情数据作为实验样本，根据提取的关键词总结群体观点语义内涵，最后对模型进行主题凝聚性测度以验证模型的观点提取能力。通过对实证结果的讨论发现：①本文以生猪期货 2022 年网络舆情数据作为研究样本，对网络舆情群体观点提取模型的效果进行了验证，结果表明，该模型能够根据网络舆情数据的语义内涵对群体观点进行合理划分；②在对网络舆情划分群体观点的同时，该模型能够按照重要性给出关键词排序，排序后的关键词对于每个群体观点的实际内涵具有较好的解释效力；③结合各个群体观点的关键词，金融监管机构能够快速、充分地了解特定时间段内期现货市场中网络舆情的演化态势，进而进行精准引导。

本文在研究中同样存在一定局限性：一是对部分主题不明确的网络信息的观点划分不够准确，部分繁杂信息会被归为一类，提取的关键词不能较好地反映该类群体观点的语义内涵；二是本文方法适用于网络舆情的文字内容，缺少对多媒体内容变化的判断。因此在下一阶段工作中，将对网络舆情群体观点提取模型做进一步改进，以更精确地识别各类群体观点，并且尝试将数据集中多媒体内容纳入网络舆情群体观点提取模型的划分过程中。

参 考 文 献

- [1] 李然, 林政, 林海伦, 等. 文本情绪分析综述 [J]. 计算机研究与发展, 2018, 55 (1): 30-52.
- [2] 吴思思, 马静. 基于感知融合的多任务多模态情感分析模型 [J/OL]. 数据分析与知识发现: 1-15 [2023-02-06]. <http://kns.cnki.net/kcms/detail/10.1478.G2.20221223.1401.001.html>.
- [3] 王萍, 李瑛寅, 郭茹燕, 等. 面向医疗文本信息的方面级情感分析 [J]. 武汉大学学报 (理学版), 2023, 69 (1): 60-68.
- [4] 睢国钦, 那日萨, 彭振. 基于深度学习和 CRFs 的产品评论观点抽取方法 [J]. 情报杂志, 2019, 38 (5): 177-185.

(下转第 177 页)

- Count Prediction Model with Paper Metadata Semantic Features [J]. *Scientometrics*, 2021, 126 (8): 6803–6823.
- [12] 任海英, 王德营, 等. 主题词组合新颖性与论文学术影响力的关系研究 [J]. *图书情报工作*, 2017, (9): 87–93.
- [13] Bosquet C, Combes P. Are Academics Who Publish More Also More Cited? Individual Determinants of Publication and Citation Records [J]. *Scientometrics*, 2013, 97: 831–857.
- [14] Weis J W, Jacobson J M. Learning on Knowledge Graph Dynamics Provides an Early Warning of Impactful Research [J]. *Nature Biotechnology*, 2021.
- [15] Abrishami A, Aliakbary S. Predicting Citation Counts Based on Deep Neural Network Learning Techniques [J]. *Journal of Informetrics*, 2019, 13: 485–499.
- [16] 赵蓉英, 郭凤娇, 谭洁. 基于 Altmetrics 的学术论文影响力评价研究——以汉语言文学学科为例 [J]. *中国图书馆学院*, 2016, 42 (1): 96–108.
- [17] Ruan X, Zhu Y, Li J, et al. Predicting the Citation Counts of Individual Papers Via a BP Neural Network [J]. *Journal of Informetrics*, 2020, 14 (3): 101039.
- [18] Wang M, Yu G, An S, et al. Discovery of Factors Influencing Citation Impact Based on a Soft Fuzzy Rough Set Model [J]. *Scientometrics*, 2012, 93: 635–644.
- [19] Rumelhart D E, Hinton G E, Williams R J. Learning Representations By Back-propagating Errors [J]. *Nature*, 1986, 323 (6088): 533–536.
- [20] Wong T C, Chan A H S. A Neural Network-based Methodology of Quantifying the Association Between the Design Variables and the Users' Performances [J]. *International Journal of Production Research*, 2015, 53 (13): 4050–4067.
- [21] Lee S, Choeh J Y. Predicting the Helpfulness of Online Reviews Using Multilayer Perceptron Neural Networks [J]. *Expert Systems with Applications*, 2014, 41 (6): 3041–3046.
- [22] 贺婉莹. 基于机器学习的论文学术创新力评价研究 [D]. 南京: 南京大学, 2019.
- [23] 胡译文, 周西姬. 基于 BP 神经网络和 MIV 算法的高价值专利预测与影响因素分析 [J/OL]. *信息资源管理学报*. <https://link.cnki.net/urlid/42.1812.g2.20231025.2230.004>.

(责任编辑: 郭沫含)

(上接第 159 页)

- [5] Fresneda J E, Gefen D. Gazing at the Stars is Not Enough, Look at the Specific Word Entropy, Too! [J]. *Information & Management*, 2020, 57 (8): 103388.
- [6] 卢国强, 黄微, 刘毅洲. 群体极化视域下突发事件网络舆情极端观点识别研究 [J]. *情报资料工作*, 2023, 44 (1): 42–51.
- [7] Yang C, Gao B, Yang J. Option Pricing Model with Sentiment [J]. *Review of Derivatives Research*, 2016, 19 (2): 147–164.
- [8] Smales L A. Asymmetric Volatility Response to News Sentiment in Gold Futures [J]. *Journal of International Financial Markets, Institutions & Money*, 2015, 34: 161–172.
- [9] Gong X, Lin B. The Incremental Information Content of Investor Fear Gauge for Volatility Forecasting in the Crude Oil Futures Market [J]. *Energy Economics*, 2018, 74: 370–386.
- [10] Lin C B, Chou R K, Wang G H K. Investor Sentiment and Price Discovery: Evidence from the Pricing Dynamics Between the Futures and Spot Markets [J]. *Journal of Banking & Finance*, 2018, 90: 17–31.
- [11] Seo S W, Kim J S. The Information Content of Option-implied Information for Volatility Forecasting with Investor Sentiment [J]. *Journal of Banking & Finance*, 2015, 50: 106–120.
- [12] Blei D M, Ng A Y, Jordan M I, et al. Latent Dirichlet Allocation [J]. *Journal of Machine Learning Research*, 2003, 3 (4/5): 993–1022.
- [13] Gao Y B, Yang L Y, Hu W J, et al. Research on Domanial Microblog Topic Evolution Based on HDP Model [J]. *Computer Engineering*, 2018, 5: 132–145.
- [14] 刘熠. 微博舆情信息受众的参与行为拟合与靶向引导研究 [D]. 长春: 吉林大学, 2020.
- [15] 庞庆华, 董显蔚, 周斌, 等. 基于情感分析与 TextRank 的负面在线评论关键词抽取 [J]. *情报科学*, 2022, 40 (5): 111–117.
- [16] 陈同辉, 鞠荣华. 中国商品期货市场价格发现能力及影响因素研究 [J]. *价格理论与实践*, 2021, (8): 118–122.
- [17] Wang S, Chen Z Y, Liu B. Mining Aspect-Specific Opinion Using a Holistic Lifelong Topic Model [C] // *International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2016.
- [18] Shams M, Baraani-Dastjerdi A. Enriched LDA (ELDA): Combination of Latent Dirichlet Allocation with Word Cooccurrence Analysis for Aspect Extraction [J]. *Expert Systems with Applications*, 2017, 80: 136–146.
- [19] Mimno D M, Wallach H M, Talley E M, et al. Optimizing Semantic Coherence in Topic Models [C] // *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, EMNLP 2011, 27–31 July 2011, John McIntyre Conference Centre, Edinburgh, UK, A meeting of SIGDAT, a Special Interest Group of the ACL. Association for Computational Linguistics, 2011.

(责任编辑: 郭沫含)