期货市场智能舆情分析与监管应用*

上海期货信息技术有限公司 支晓繁 薛利 支文纲

一、研究背景及创新点

（一）研究背景

近年来，以大数据、云计算和人工智能为代表的金融科技的快速发展为期货市场带来全新的技术变革。金融科技在带来金融效率提升和市场快速发展的同时，也对期货市场的监管工作提出了新的挑战。作为期货市场重要的基础设施，期货交易所需打造数字化和智能化的监管体系，不断提升监管效能，对期货市场上发生的异常事件进行及时分析、处理，维持期货市场平稳、有序地运转。

鉴于互联网时代舆情信息呈现出扩散快、影响面广、传播链条复杂等特点，互联网舆情对期货市场的影响越来越不容忽视。作为对信息高度敏感的行业，期货市场上存在的海量舆情蕴含了大量有价值的新闻线索和价格趋势。与此同时，社会化媒体产生的数据具有规模大、异构多元、组织结构松散的特性，对舆情的全面采集、分析和处理提出了更高的要求。特别地，舆情分析需要对可能影响市场运行的负面舆情进行及时预警，以支持监管部门全面准确地把握舆情动态，实施针对性监管。如何借助新技术手段形成智能化的舆情分析能力，实现对海量舆情的高效搜索和相关实体间的关联性分析，是监管部门需要着力考虑的问题。在此背景下，拟建设期货市场智能舆情分析与监管应用，旨在对舆情信息开展语义分析，提取舆情中的非结构化信息，在此基础上结合相应的舆情数据进行复杂关联分析，以满足多种典型监管场景对智能舆情应用的需求。

（二）创新点

本文遵循“应用导向、业务导向、监管导向”的原则，实现期货市场智能舆情分析与监管应用。在实现舆情数据获取、舆情数据抽取、舆情数据清洗、舆情数据预处理等基本处理及分析功能基础上，聚焦期货市场监管业务痛点，研究利用大数据和人工智能技术，实现面向典型监管场景的舆情分析应用功能，围绕事件趋势判断、市场风险的预测与推演、舆情线索发现等应用场景开展了针对性的舆情分析研究。通过构建围绕热点事件相关的知识图谱，针对主体关联度及传播路径的分析，同时排查主体画像的司法风险、交割风险、信用风险等风险指标，实现了对事件趋势判断及市场风险的研究，进一步完善了“热点事件分析”与“风险排查”业务分析闭环。具体而言，主要包括以下几个方面：一是引入自研光学字符识别（Optical Character Recognition，OCR）、语音识别、视频识别等新技术，实现了对文本、图片、视频等多源数据的处理及整合；二是通过风险传播路径触达主体的风险分析，量化

* 本文为 2019–2020 年度上海期货交易所优秀研究成果。

事件中相关主体的风险等级、影响趋势，为防范危险提供了风险预警与事前干预的有力依据；三是基于自然语言处理 (Natural Language Processing，NLP) 模型对舆情数据进行实体提取及预分类，同时结合知识图谱（Knowledge Graph，KG）技术实现舆情风险的提前预警；四是在实践中充分考虑了技术的服务化需求，将多种技术封装为标准化的细粒度服务，可灵活支持不同应用场景的分析需求，有效提升相关技术服务的泛化推广能力，为行业智能舆情分析技术落地提供了较好的实践经验。总体来看，本文的研究成果兼顾技术创新性和应用实效性，已取得成果并且在研究深度和业务融合方面具有较强特色。

二、智能舆情分析基础能力建设

从技术能力建设角度看，期货行业智能舆情分析能力主要体现在舆情信息基础处理、行业场景化舆情分析，以及舆情可视化等三个方面。其中，舆情信息基础处理能力主要指舆情数据获取、舆情数据规范化、舆情数据清洗、舆情数据预处理等基础处理能力；行业场景化舆情分析主要围绕期货市场的典型业务场景，例如事件风险传导，网络黑嘴识别等，开展基于智能舆情分析的创新应用；舆情可视化能力主要借助大数据可视化技术，尤其是对非结构化数据和多维度结构化数据的展示技术进行业务视图优化，提升用户体验。

（一）期货行业舆情基础处理

智能舆情分析的基础服务功能主要包括：期货舆情基础数据库构建、舆情数据采集、舆情预处理、舆情要素获取、舆情要素处理、舆情标签化处理等功能。

期货行业舆情基础处理是指利用自然语言处理技术对来自于互联网、移动互联网，尤其是“两微一端”、证券期货类资讯网站、知名网络论坛等的舆情数据进行自动化抽取、过滤、清理、识别及整合，从中提取有价值的信息，为具体业务提供外部舆情支持。构建舆情基础处理能力是实现期货行业场景化分析应用的基础任务之一，现已在行业内形成一定共识，其具体内容包括行业基础数据库构建、舆情数据采集、舆情预处理、舆情要素获取、舆情要素处理、舆情标签化处理等技术能力。

1. 行业基础数据库的构建

行业基础数据库的构建主要为舆情分析处理提供数据存储，整合及管理的手段，内容包括基础舆情、实体库、金融词库、期货品种知识库以及标签库的构建，具体可以通过专家协助、众包以及人工智能学习算法来逐步构建和完善，其中部分关键内容如下：

(1) 实体库：实体对象主要涵盖期货行业产业链上下游重要的生产企业、流通企业、消费企业以及相关的期货公司，仓库、物流、银行以及其他间接参与期货交易的金融机构等。

(2) 金融词库：围绕期货行业、期货品种、期货术语、宏观经济、产业链关系等进行定义。

(3) 舆情标签：主要根据舆情分析所要解决的业务场景进行信息扩展，通过专家分类将不同场景和舆情标签进行关联，以辅助后续业务场景下对特定舆情集合分析的需求。

2. 舆情数据采集

舆情数据采集主要以实时和批量方式从外部舆

情数据源获得基础舆情数据，并根据采集时效性和数据源覆盖度、权威度进行效果评估。其中，实时舆情主要指通过预先配置的监控规则对目标采集源进行变化地持续跟踪，当采集源中发生变化后，变更消息将推送到采集任务，并触发采集任务。批量数据导入指采集通过第三方舆情服务商提供的数据接口定期批量导入舆情数据。按照数据采集手段，可分为如下三种方式：

(1) 网络爬虫：从新闻、论坛、博客、平媒等传统网络媒体，微博、微信等新媒体以及海外媒体爬取相关期货舆情数据；

(2) 元搜索：对各类搜索引擎，例如百度、谷歌、360、搜狗等进行定向内容的采集聚合；

(3) API 接口：通过网关与第三方舆情数据服务商进行连接，采集相关舆情信息。

3. 舆情预处理

舆情预处理主要指对原始舆情数据的结构化处理及数据清洗，旨在为后续数据分析提供统一的结构和质量保证，主要手段包括超链分析、编码识别、URL 去重、锚文本处理、垃圾信息过滤、关键字抽取、关键信息提取、正文抽取、自动摘要、非结构化内容识别和转换等方式。

4. 舆情要素获取

舆情要素获取主要指对原始舆情文本进行内容要素提取和抽取，并进行统计和建模分析，其基本流程如下（图 1）：

图 1：舆情要素获取基础流程图

(1) 舆情元数据提取：包括标题、内容、舆情来源、涉及关键词、首发时间、转发次数、转发媒体、转发时间、评论数量等。

(2) 文本预处理：包括文本分词、词性标注、去除停用词、命名实体识别等。

(3) 文本统计：包括词频统计、语义向量计算、关键词识别、语义网络分析等。

(4) 深度信息提取：主要包括主体关系识别、属性提取、主题汇总、关系抽取等。

在上述提取过程中，需要通过知识融合的方法对实体进行识别，消除实体描述歧义，建立对真实实体的链接关系，并以此形成实体关系网络。

5. 舆情要素处理

舆情要素处理主要指支持复杂舆情分析应用而开展的个性化舆情要素统计、建模和分析，根据分析对象级别可分为单篇和多篇舆情要素统计。此类任务一般与具体业务场景紧密联系，基本处理流程如下（图 2）：

(1) 单篇舆情要素处理：主要根据文本内容在整个舆情文本库中出现的频度，对单个舆情的核心内容进行识别，提取分类信息，寻找内容相似性以及对实体情感的关系等。

(2) 多篇舆情要素处理：指在单篇舆情要素分析基础上，进行多篇文本间的要素关联分析，具体可

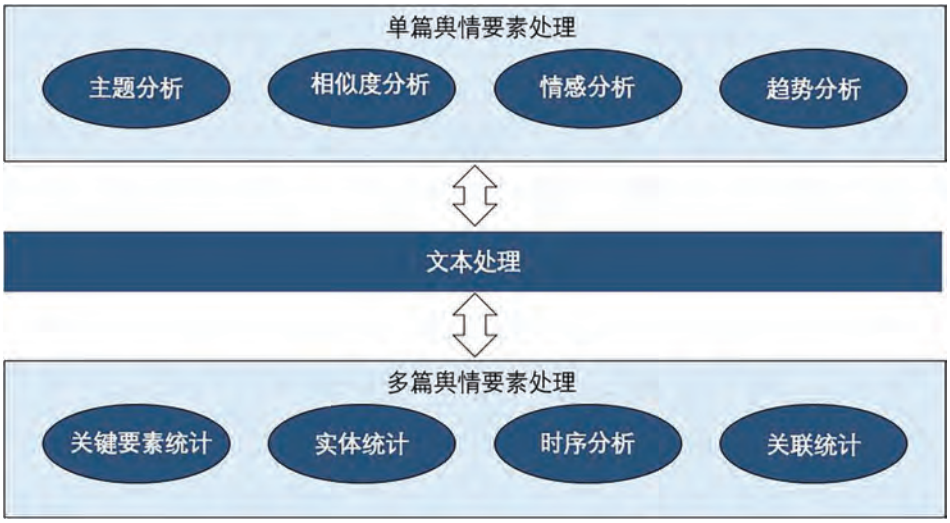

图 2：基础流程图

细分为同一主题的多篇舆情要素、同一主题时序化的多篇舆情要素、同一实体的多篇舆情要素、同一实体时序化的多篇舆情要素四类。

6. 舆情标签化

舆情标签化是在舆情基本处理操作基础上，通过对舆情信息特征进行归纳抽象，实现舆情元素的“再加工”。按照标签的生产方式可分为人工方式，或基于规则、算法模型的自动化方式。与人工方式相比，基于算法模型的标签生产方式一方面可提供多角度的标签逻辑，同时也可根据实际情况实现标签体系的自动化更新。此外，标签关系的识别是标签化处理的另一个重要内容，通过建立多个标签的关系结构，比如将风险标签与市场波动标签建立关联，可进一步发掘舆情信息的隐含属性，从而实现舆情关联分析。

（二）期货行业场景化舆情分析

作为行业舆情分析的关键能力之一，场景化分析是以情感分析、主题分析等技术为支撑，将舆情分析与现有各类业务场景结合，实现场景化的舆情服务。例如借助智能化的舆情情感分析技术，实现对市场情绪或重大事件的跟踪，可针对全市场以及品种的走向进行预测，帮助监管人员及早发现相关异常行情；亦或利用知识图谱技术，实现市场风险传导分析等。

（三）期货行业智能舆情分析应用设计

结合前文介绍的舆情分析能力建设的思路，期货行业智能舆情分析可按照三层功能架构进行设计，具体如下（图 3）：

1. 基础数据层

该层主要完成数据采集、存储和处理的任务。提供对接口区多源数据的清洗、整合、加工和存储，为核心存储服务。考虑访问的一致性，核心层中建立的非结构化数据主要通过 Hadoop HDFS 以目录文件方式进行存储，并通过建立索引表的方式对非结构化数据提供 SQL 检索访问。

具体而言，数据采集可采用 Sqoop 工具实现批

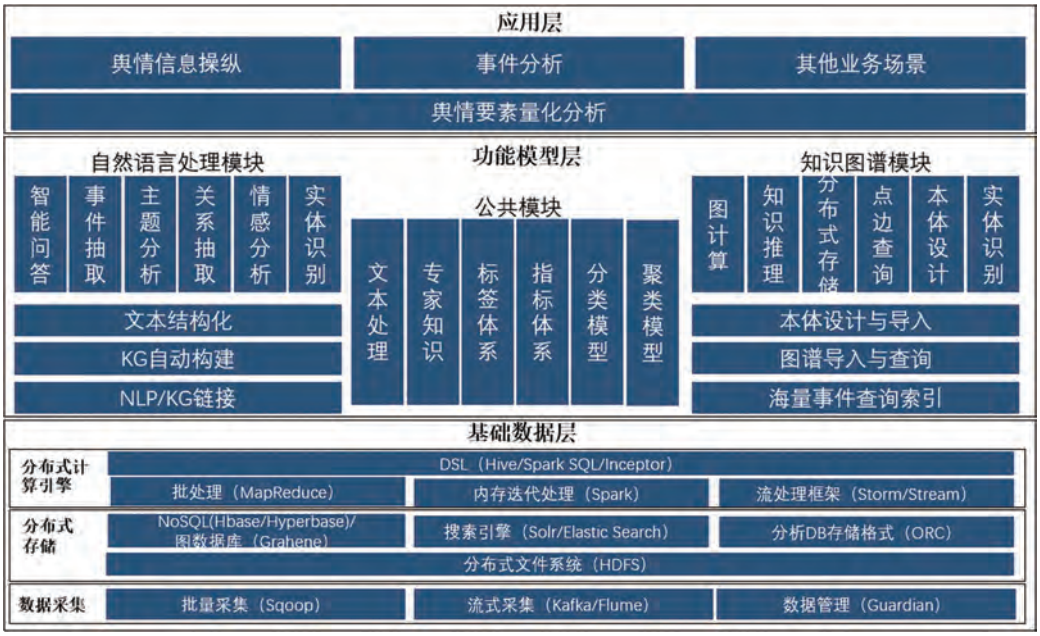

图 3：智能舆情分析平台逻辑功能架构

量采集，Kafka 或者 Flume 进行流式采集；此外，舆情数据存储主要考虑分布式存储，底层为统一分布式存储 HDFS，HDFS 采用三份副本策略保证数据的安全性以及可靠性。在 HDFS 之上提供分布式 NOSQL（Not Only SQL）实时数据库 Hyperbase，为高并发检索分析提供平台级的支撑。支持各类结构化、半结构化、非结构化海量数据的低成本存储，为超长时间的海量历史数据存储和使用提供基础支撑。而数据处理主要在分布式计算引擎基础上，通过 YARN 提供统一的资源管理调度，并借助 Spark 计算框架的并行统计算法库和机器学习算法库，为舆情处理平台提供高效的数据挖掘能力。同时，基于传统的 MapReduce 计算框架，实现对各种大数据计算框架的支持。此外，可利用 Stream 实现对实时数据的低延时高吞吐的处理，并根据组合消息队列 Kafka 技术，打造适用于各种实时数据的复杂处理场景。

2. 功能模型层

该层主要为舆情处理分析提供算法模型支持，旨在实现知识的结构化。该层设计的一个重要关注点是尽量提高主流算法模型的覆盖度。功能模块一般包括自然语言处理模块、知识图谱模块。自然语言处理模块主要支持实体识别、关系抽取、情感分析、主题分析、事件抽取和智能问答、文本结构化、KG 自动构建和自然语言处理、KG 链接处理。知识图谱模块则需提供本体设计、点边查询、分布式存储、图计算、KG 本体设计与导入、KG 图谱导入与查询和 KG 海量事件查询索引等功能。

3. 应用层

应用层主要结合业务的场景化需求进行个性化设计，功能上尽量覆盖常见的业务场景。以异常舆情事件分析场景为例，异常舆情事件相关要素量化

任务一般会涉及事件影响因素量化、媒体统计量化、情感量化等需求，因此该场景的舆情分析服务需覆盖上述功能，此外可视化设计方面还需结合事件量化的各种指标特点进行针对性设计（图 4）。

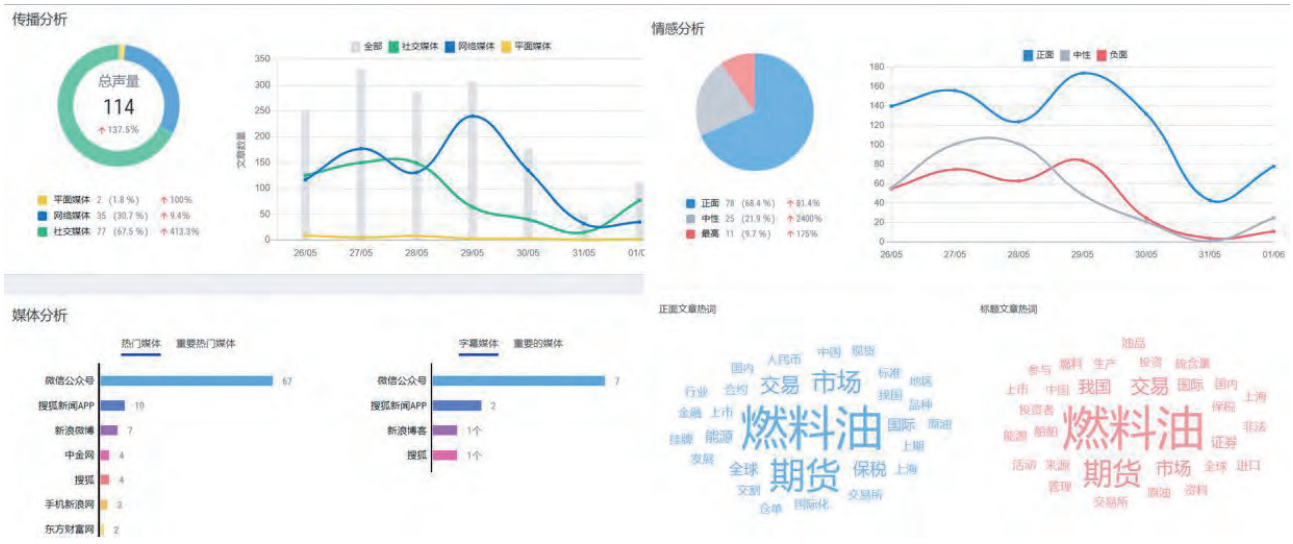

图 4：舆情要素量化可视化

（四）模型研究成果

1. 实体抽取模型

实体抽取有助于舆情的深度分析。目前主流的实体抽取模型主要包括 Bi-LSTM+CRF、Lattice-LSTM+CRF、Join Model 和 Matching the Blanks（MTB）。Bi-LSTM+CRF 是一种经典的运用深度学习完成序列标注任务的模型，在英文语料中取得了较好的效果，其结构图如图 5 所示。Lattice-LSTM+CRF 模型增加了对字粒度、词粒度的考虑，相对于只使用字粒度的模型，增加了词信息，丰富了语义表达，相对于仅使用词粒度的模型，可以避免分词错误带来的影响，其结构图如图 6 所示。Join Model 基于联合抽取的思想，通过一个模型实现端到端的实体识别和关系确认，其结构图如图 7 所示。尽管三种模型在实体抽取中取得了一定的成就，但是它们都需要预定义实体关系类型，再识别实体之间是否存在该种关系，因此无法挖掘新的实体关系类型。Matching the Blanks（MTB）设计一个通用的关系抽取器，提高了泛化能力，无需事先定义好关系，可以对任意关系进行抽取建模，其结构图如图 8 所示。

Lattice-LSTM+CRF 和 Join Model 模型都需要预先定义好关系类型，然后识别实体之间是否存在该种关系，因此缺乏发现新关系类型的能力，而模型 MTB 可以很好的补充了对新关系的发现。因此，本文提出了一种基于混合模型的多级实体关系抽取策略，首先利用 Lattice-LSTM+CRF 和 Join Model 对可以预定义、有规则的实体关系进行抽取，同时

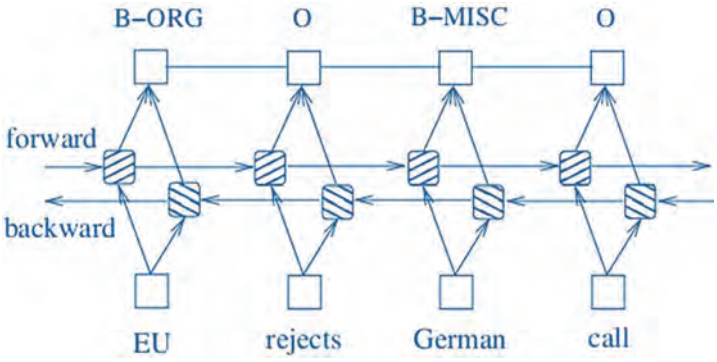

图 5：Bi-LSTM+CRF 模型结构图

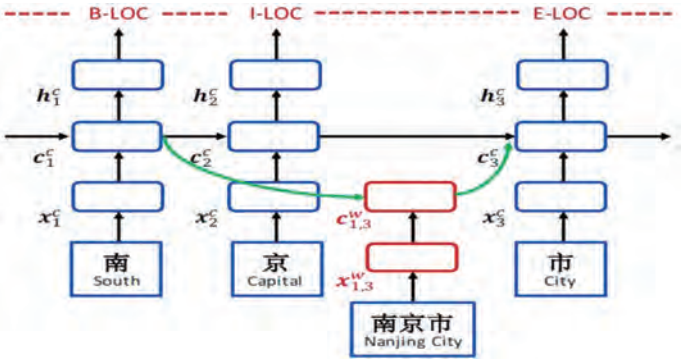

图 6：Bi-LSTM+CRF 结构图

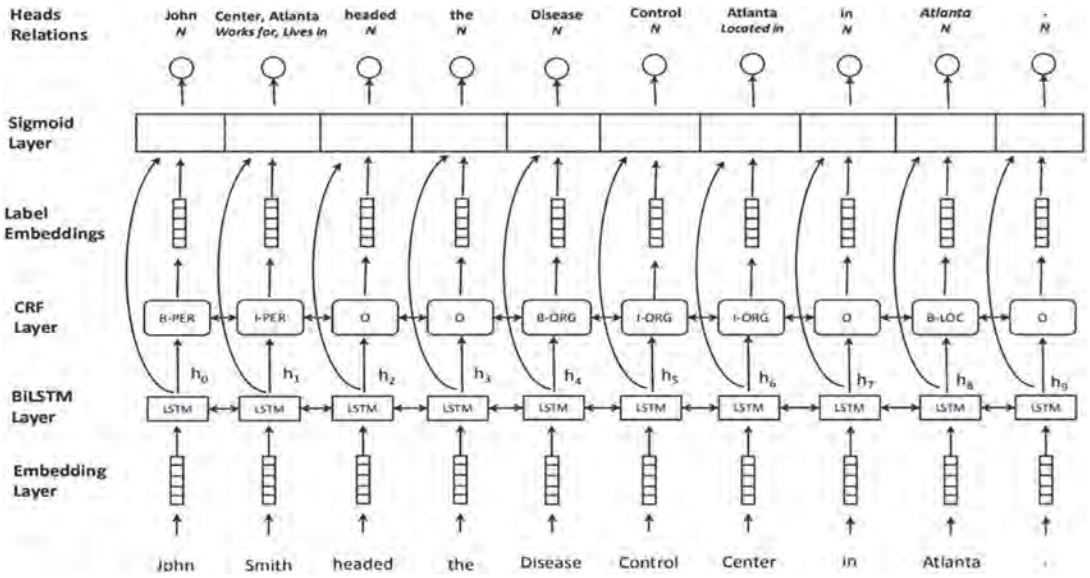

图 7：Join Model 结构图

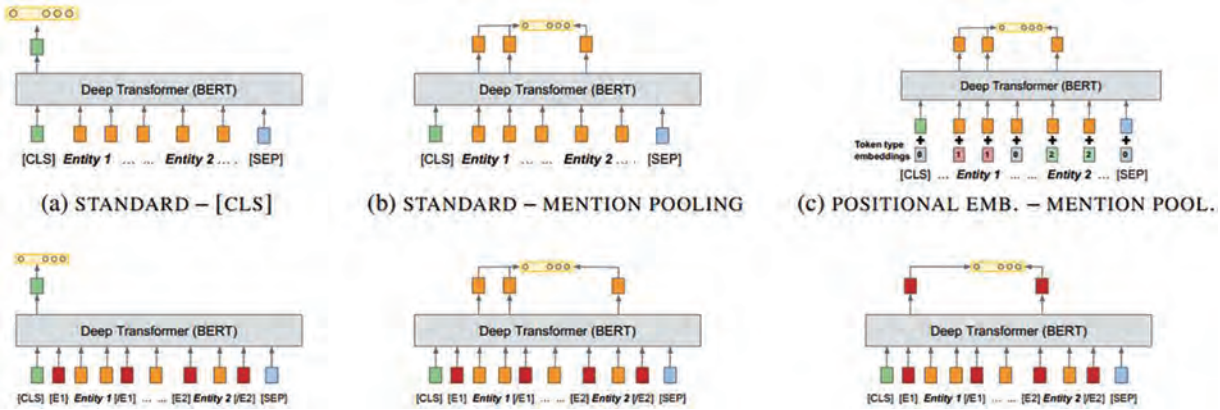

图 8：MTB 结构图

利用 MTB 模型挖掘新关系以进行补充，实现对实体关系类型的初步抽取，随后将抽取的实体关系类型输入到 Bert 模型中，以实现实体关系类型的深度挖掘。混合模型的多级实体关系抽取策略的实现思路如图 9 所示。

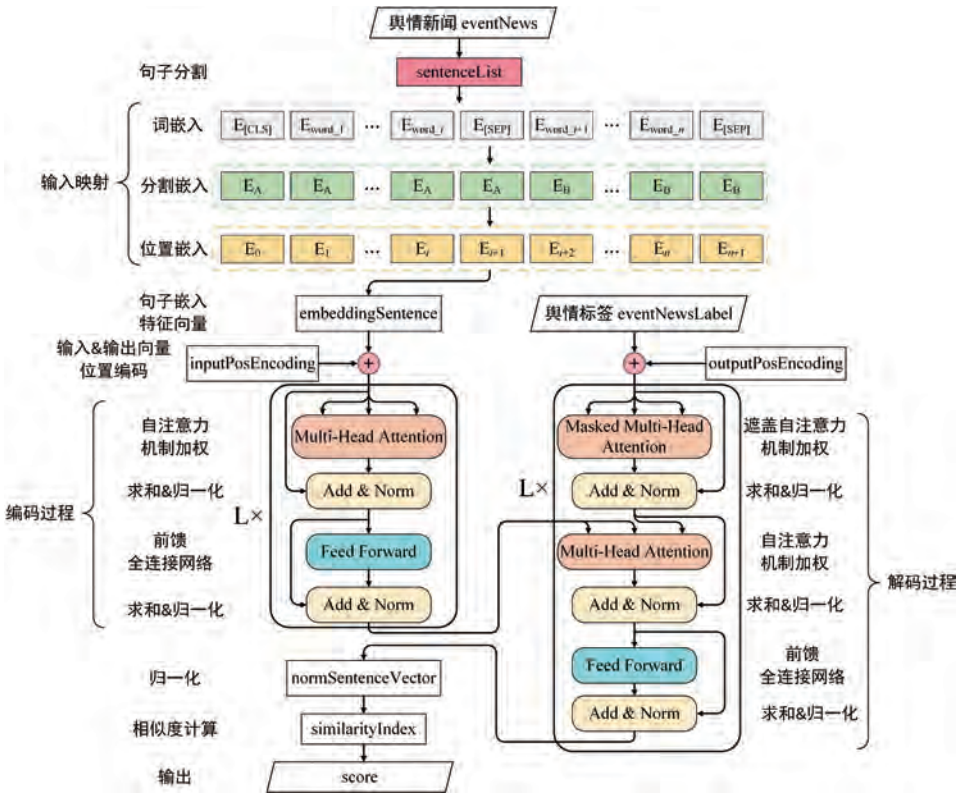

图 9：实体抽取模型示意图

2. 混合文本识别模型

随着社交媒体的发展，舆情不再局限于以文本的形式传播，图片等非结构化数据同样可能蕴含着重要的信息。非结构化文本数据的自动化字符识别有利于助推监管机构提升监管效能。电子化文档可以根据格式分为已知模板和未知模板的文档。对于已知模板的电子化文档而言，有效文本未知相对固定，传统文本识别方法缺乏对已知信息充分利用的能力，导致识别效率低下；对于未知模板的电子化文档而言，由于文档背景相对复杂，文本位置不固定，需要尽可能保留所有文本信息。

因此，本文提出了一种全新的混合文本识别模型。在文本定位阶段，对于已知模板的电子化文档，首先利用人工标注锚点在模板库中自动匹配模板，并获取该模板下预标注的有效文本区域；若文档中印有红章且覆盖文本区域，则采用基于颜色通道的快速行进算法进行图像修复，再执行文本定位工作，进而获取标准化的文本行区域。对于未知模板的电子化文档，为了尽可能捕获所有文本特征，利用 CTPN 网络扫描整个文件，将提取的特征集成到双向 LSTM 中以捕获长距离依赖关系，并通过全连接网络定位框获得最终所有文本结果。在文本识别阶段，全部采用由 CNN、双向 LSTM 和 CTC-loss 组成的 CRNN 模型，实现变长文本识别并输出最终的文本识别结果（图 10）。

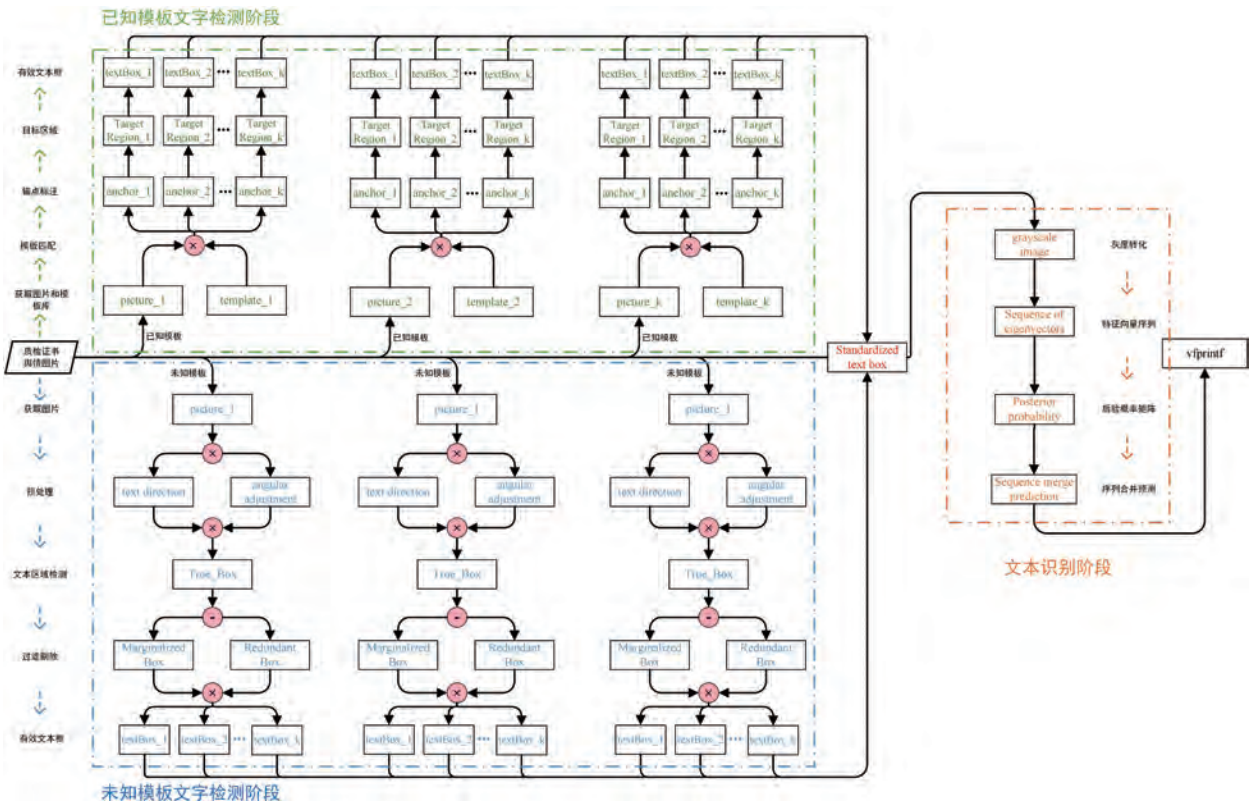

图 10：混合文本识别模型示意图

非结构化文本数据的自动字符识别能够有效提升单证核验工作的效率，从而帮助监管机构监督商品期货交易行为。对于有固定模板的单证文档而言，传统文本识别方法缺乏对已知信息充分利用的能力，导致检测效率低下；对于无固定模板的文档而言，由于文档背景相对复杂，文本位置不固定，需要尽可能保留所有文本信息。因此本论文提出了一种全新的面向单证核验混合文本识别方法。在文本定位阶段，对于存在固定模板的单证，首先利用人工标注锚点在模板库中自动匹配模板，并获取该模板下预标注的有效文本区域；若单证中印有红章且覆盖文本区域，则采用基于颜色通道的快速行进算法进行图像修复，再执行文本定位工作，进而获取标准化的文本行区域。对于未知模板的单证，为了尽可能捕获所有文本特征，利用 CTPN 网络扫描整个文件，将提取的特征集成到双向 LSTM 中以捕获长距离依赖关系，并通过全连接网络定位框获得最终所有文本结果。在文本识别阶段，全部采用由 CNN、双向 LSTM 和 CTC-loss 组成的 CRNN 模型，实现变长文本识别并输出最终的文本识别结果。本文提出的混合识别方法可应用于质量检验证书和商业合同的文本识别，实现了可靠的文本识别结果，并大大节省了人工成本，提升了检测和识别效率。

3. 舆情线索识别模型

财经新闻是投资者进行研究和投资决策不可缺少的来源。然而，也有许多虚假的财经新闻涌入人们的日常生活。这类信息可能影响舆论，为一些犯罪分子操纵金融市场提供机会。随着直播、短视频等行业的兴起，除了传统的文本或者图片的形式外，视频、音频等多媒体中也包含海量的重要信息。此外，社交媒体中的评论、弹幕等信息同样可以为虚假新闻识别提供支撑。面对复杂、多元的新媒体新闻，传统的虚假新闻识别方法略显单一，局限性大。此外，现有的虚假新闻模型大多采用暴力挖掘的特征，对于部分难以察觉的特征无能为力。

因此，本文提出了一种基于多元化数据特征决策支持的虚假新闻识别，针对不同的新闻载体采用不同的方式。首先，利用视频识别、语音识别、文字识别等技术获取多元新闻数据的文本信息，同时结合用户评论、新闻来源及市场数据，深度挖掘多元化文本数据中的观点特征、情感特征、用户特征及传播特征，以实现多元数据下的虚假新闻检测。在构建特征时，选用提取抽象特征能力更强的神经网络模型。模型整体分为嵌入层、编码层、交互层、解码层以及输出层。嵌入层负责将文本输入转化为机器可以处理的稠密向量和字符级别的信息，编码层使用深度神经网络提取高维度特征，交互层使用注意力机制帮助模型学习各输入之间的关系，解码层使用深度神经网络将特征解码成输出字符的概率分布，输出层则直接输出对应的预测结果。多元化数据特征决策支持的虚假新闻识别的特征提取示意图如下（图 11）。

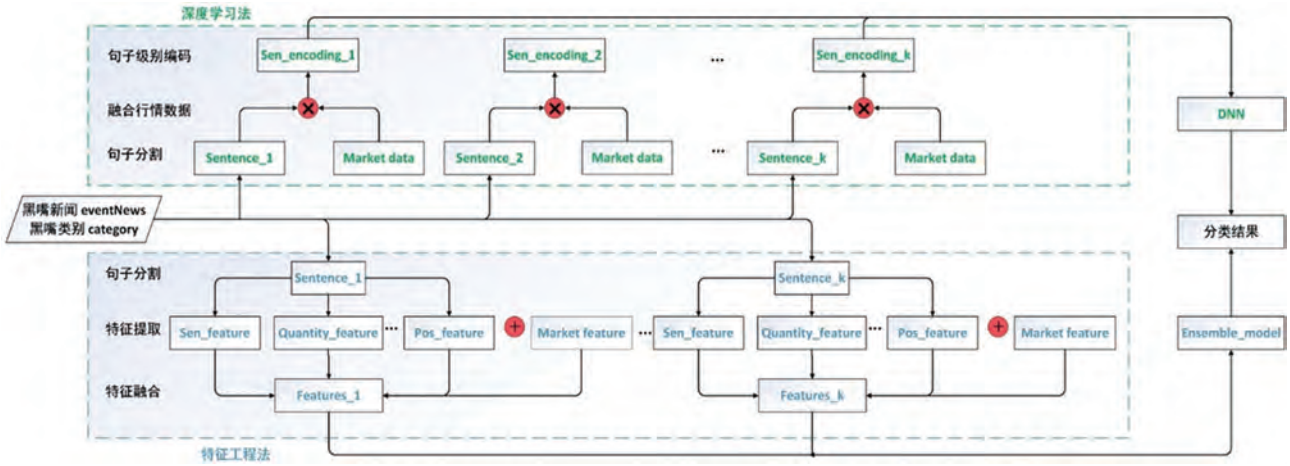

图 11：舆情线索识别模型示意图

4. 趋势判断模型

传统的针对金融行业趋势判断的模型，往往都是通过“关键词 + 规则”进行判断，针对各实体之间的关系，通常采用以本体概念为基础的知识图谱形式，这种形式存在一些局限性，例如实体语义不完善，容易断章取义，不能有效地传递出各实体之间的逻辑关系等。

因此，本文提出了一种基于专家知识推理的事理逻辑判断趋势模型。该模型首先利用基于 TF-IDF 和 TextRank 的模型寻找原始文本关键词语实体，利用 MMR 和权重过滤相似实体并连接合并得到重要目标实体短语。其次，利用基本句法关系寻找句子之间的六种逻辑要素，包括：因果、转折、条件、顺承、并列、反因果。接着，利用基于哈工大的 PYLTP 模型，对整理的六种逻辑要素句子对进行语义句法分析，寻找重要目标实体的前后谓词性事件并连接得到谓词性事件实体。最后，利用情感分析或者专家知识库规则匹配得到谓词性事件重要实体对品种的趋势影响，并根据不同场景需要，选择是否形成图形式的关系网络图。

基于专家知识推理的事理逻辑判断趋势模型能够解决谓词性事件实体之间、谓词性事件实体之间状态的逻辑关系，保证了重要目标实体语义完整性和谓词性事件实体状态之间的清晰性，从而更准确地判断舆情类新闻对品种实体的影响趋势。基于专家知识推理的事理逻辑判断趋势模型的示意图如下（图 12）。

5. 基于知识图谱的风险传导模型

基于知识图谱的传导模型构建流程如下：

(1) 事件及风险定义

事件发现分为两类：一是通过舆情标签、指标规则定义事件。二是通过前端指标的自定义设置。

(2) 事件脉络分析

根据事件的定义规则查询大数据平台，获得同类事件的历史数据，进行事件脉络分析，包括事件时序分析、事件发展走势预测等。

(3) 事件专题分析

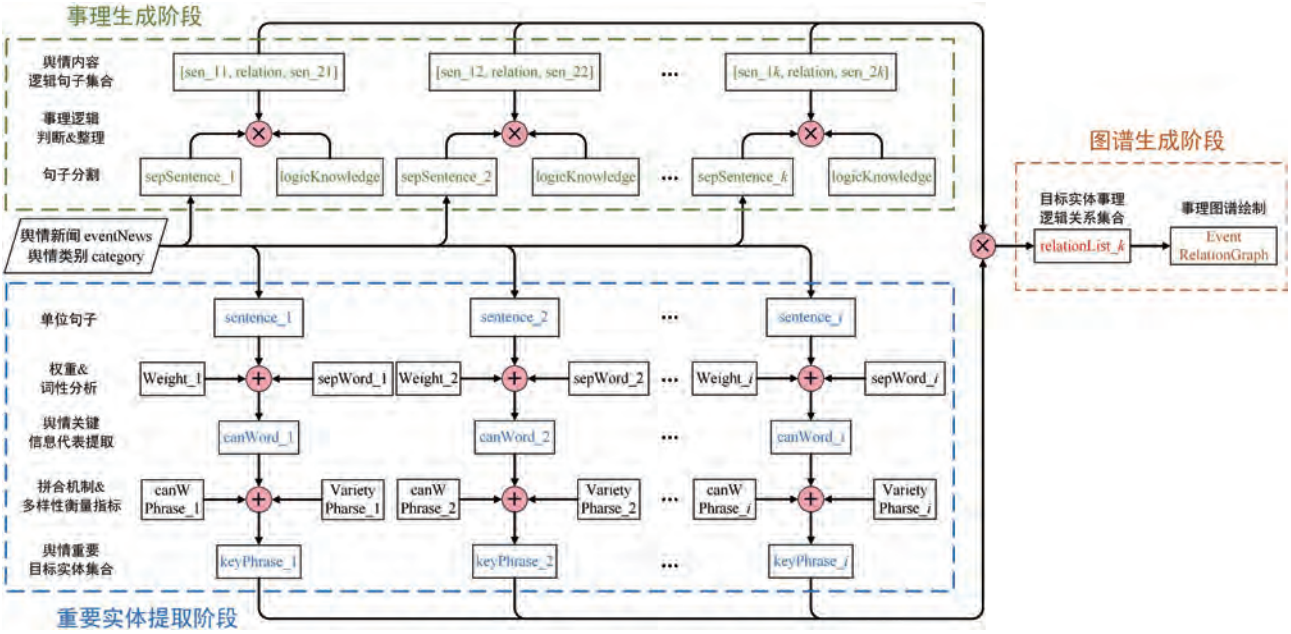

图 12：趋势判断模型示意图

对事件的情感预判，对事件的主题分类，对市场的影响进行趋势分析，结果同步历史事件表。

(4) 事件传导分析

根据事件对市场的影响趋势预测对事件的传导进行分析，并高亮在行业知识图谱的传导路径。

(5) 事件风险分析

对事件中涉及的实体进行风险分析，综合评估事件的风险，识别潜在风险。

三、应用场景成果

（一）OCR 数据识别

1. 非结构化文识别

使用通用文字识别技术，实现对各类文档、表格、财报、票据等纸质文档的识别，并返回文字在图片中的位置信息以便于进行比对、结构化等处理，可满足各个业务部门的文档快速录入、存档和检索的需求，有效降低人力成本，提高信息录入效率。

2. 基于信息抽取的关键信息冲突检测

基于实体信息冲突检测的参与者报送信息核验的思路主要包括利用非结构化信息抽取或 OCR 技术的实体信息提取、实体信息与信息库数据冲突检测，以及基于冲突检测的异常信息预警等关键步骤。以 OCR 仓单识别场景为例，利用 OCR 技术仓单 PDF 数据的实体信息提取对仓单的图片信息和相关的 PDF 中的基本信息进行提取，将非结构化数据转换为结构化数据，识别、提取出涉及的关键信息，示例如下（图 13）。

图 13：基于信息抽取的关键信息冲突检测应用场景

报送数据冲突检测是指从非结构化的报送数据中抽取相关信息，与已存在的信息进行匹配验证，对检测出的数据冲突予以提示。例如报送的仓单信息与主体信息库中注册信息的核验。交割过程中，业务人员需要对大量的仓单数据进行核查复检，其中大部分的数据没有问题，但是往往耗费大量的时间，而少量的问题数据更需要人工着重核查和复检。通过利用实体信息冲突检测快速筛选出仓单信息与主体信息库中注册信息不匹配的数据，再进一步进行人工核查，可大大缩短筛选时间，提升核验效率。

（二）舆情线索分析

1. 基于规则的线索发现

通过对异常舆情的分类及违法违规行为模式的梳理，从各类舆情数据处理分析是否存在疑似违法违规行为的特征，通过构建风险词库、实体关系库、负面词库，通过词库建立标签，进一步利用标签向量聚类生成舆情的标签树，实现对舆情的异常类型、风险类型，以及文章中涉及主体做出线索提示。标签分类的舆情有助于业务人员快速发现、定位问题，提高工作效率。

2. 新媒体舆情线索的探索发现

随着 5G 网络和新媒体的高速发展，舆情新闻

的传播方式趋于多样化，相比于传统的文字载体，如今更多的舆情消息是以图片、视频等形式进行传播、展示。因此舆情图片中的文字对于监管机构获取相关金融信息，从而规范金融市场、监督交易行为而言是十分重要的。本文采用基于 PaddlePaddle 的深度学习框架搭建的舆情图片 OCR 模型。该模型适用于多种场景下舆情图片的文本检测，模型对于文字区域的检测速度快，识别结果精准，模型能够将舆情图片中结果文字相对位置与识别结果同时输出，能够有效地保留图片中所有的舆情文字，保证信息的完整性。下图为利用 PaddlePaddleOCR 模型对舆情图片检测并识别出对应结果图（图 14）。

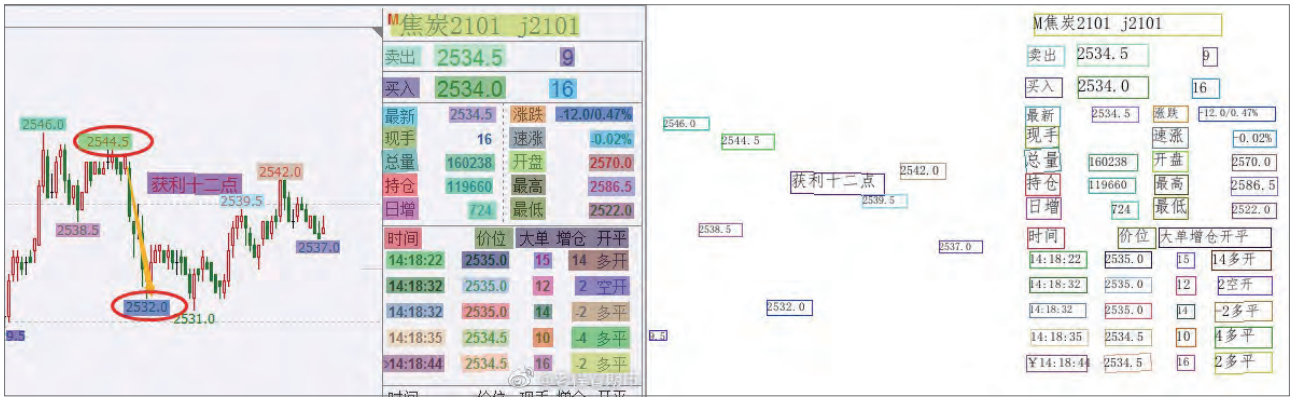

图 14：舆情图片检测并识别出对应结果图

除了舆情图片之外，短视频、直播等形式随着信息化的高速发展也逐渐兴起，部分发布者利用视频直播的形式散布虚假信息，因此虚假新闻传播形式更加多元化。针对视频、直播的视频格式也是由逐帧图片组成的特点，以及直播类视频中常常有主播、主讲人解说却不配字幕的问题，本文提出了一种基于逐帧切分和音频文字识别的舆情视频识别模型。该模型利用固定的时间窗口，对舆情视频进行图片切分，并对切分的帧图片进行 OCR 识别，识别模型采用图片识别的 PaddlePaddleOCR 模型，保证获取帧图片内文字信息的完整性；其次对视频中所有获取的文字结果进行去重，保留所有重要且不重复的舆情信息；然后利用音频识别 API，对整段视频进行语音文字识别，基于逐帧切分和音频文字识别的舆情视频识别模型识别，输出标准化结果；最后将所有结果保留，作进一步的舆情分析。该模型能够保证尽可能获取视频中的所有文本信息，包括帧图片文本以及语音文本，最大限度获取视频内所包含的所有文字结果，为后续舆情分析作充分准备。

（三）舆情事件分析

期货市场作为资本市场的重要组成部分，其运行具有对信息高度敏感的特性。市场参与者对行情趋势的判断，往往会形成集体舆情情绪及偏好；人工方式的收集及分析，覆盖率低且效率不高，难以为市场监控与分析提供有力支持。借助智能化的舆情情感分析技术，实现对市场情绪或重大事件进行

跟踪，可针对全市场以及品种的走向进行预测，帮助监管人员及早发现相关异常行情。

1. 基于分类算法的实体趋势四级预测

基于分类算法的实体趋势四级预测，主要是多个 NLP 模型的集成应用，其关键步骤如下：

(1) 实体识别: 提取舆情文本中涉及的实体信息，包括实体、关系、属性。

(2) 观点分类：针对提取的实体，通过文本中词句结构，提取出观点属性。

(3) 情感分类: 对舆情进行主题相关的情感分析，收集大规模噪声标记数据，在此类数据上通过依存句法分析提取主题相关的特征以及使用的五级情感分类主题模型。在主题相关的情感分析中，情感的对象不是舆情正文，而是文本中特定的被描述的实体。针对每个实体判断期货市场总体情感倾向的分类，可分为正面、中性、负面三类。

(4) 趋势判断：与情感分类类似，针对提取出的实体，通过利用机器学习分类算法模型对舆情数据进行分析判断期货市场针对品种的行情趋势的分类，分为利多、中性还是利空三类。

(5) 风险预测分类：基于舆情数据，对市场情绪或重大事件进行跟踪，对各单一品种或市场的走向行预测，以便监管业务人员对可能出现的异常情况提前对风险等级的划分及风险提示，为业务人员准备防范措施争取时间。

2. 基于热点事件跟踪与预警

事件标签库建模是指通过事件时间戳、事件主题、事件标的、事件标签等关键维度对某次异常波动进行事件化描述。其中，事件标的和事件标签主要用于舆情文本与涉事标的的关联分析，关键步骤如下：

(1) 异常事件的识别与检查

主要利用基于异常规则或异常识别模型的方式实现异常事件的检测识别。其中，基于异常规则的方式主要是通过预定义的异常指标体系及检测规则实现异常事件的识别；基于异常识别模型的方式主要利用行业监管科技 3.0 技术协作组《基于 Δ X/ Δ T 模型的证券市场异常交易检测方法》提出的异常事件检测方法实现异常事件识别。

(2) 基于历史事件的标签体系与知识体系的事件关联性分析

主要利用历史事件的标签体系与知识体系的对事件进行关联分析。

(3) 热点事件的多维度分析

热点事件分析包括事件观点的分类、情感的分类、重点关键词云等的多维度的对比分析，帮助业务人员分析热点事件。

(4) 事件风险预警

基于舆情数据，对市场情绪或重大事件进行跟踪，对各单一品种或市场的走向行预测，以便监管业务人员对可能出现的异常行情提前对风险进防范措施。

（四）实体风险传导判断

实体风险分析的目的是利用舆情信息对资本市场主体的各类描述或判断，对舆情中各类参与者的行为特征或重大事件进行归集、分析，预测实体可能即将爆发的风险类型及相应线索，实体风险分析可分为实体经营行为风险分析、实体诚信风险分析。

1. 基于特征归集的实体画像经营风险分析

利用舆情信息对资本市场主体的各类描述或判断，对舆情中各类参与者的行为特征或重大事件进

行归集、分析。实体经营行为风险分析是基于舆情信息对期货市场参与者的各类经营行为进行监控，识别潜在的风险。实体经营行为风险分析可通过提取单篇舆情要素 (风险关键词) 与标记风险标签的方式实现。主要思路及步骤分为如下：

(1) 实体画像

(2) 风险量比图

(3) 风险预警分析

2. 异常舆情的风险传导分析

基于知识图谱传导路径的风险传导分析，通过对舆情，对相关舆情追踪溯源，主要思路及步骤分为如下：

(1) 舆情事件图谱构建

利用舆情文本中提取的实体、关系、属性构建事件图。

(2) 风险传播路径分析

风险路径的分析根据提取出的前序、后序原因构建事件的发展路径，同时利用风险实体的属性值的规则配置。

(3) 风险推演

该步骤主要利用之前分析的结果，对关联舆情进行可能风险的推演，舆情实体要素，根据要素对应风险等级，并计算预警指标，以舆情实体和风险实体作为两端，利用产业链图谱进行传导推理，获得传导路径，找出异常事件的导火索，挖掘更深层的关联方，实现对潜在风险的挖掘与预警。

四、总结与展望

本文的核心目标是借助智能分析技术来解决期货市场典型监管场景中的业务痛点问题。从项目落地情况来看，有效地解决了相关的非结构化文本的分析需求，场景化舆情分析与应用，以及舆情分析可视化等三方面内容。

伴随着金融科技在期货市场的日渐普及，人工智能、大数据等新兴技术不仅深刻改变目前，行业监管机构不断加大行业对大数据、人工智能等新兴技术在监管工作中的研究及应用力度。就本文探讨行业智能舆情分析监管类应用而言，相关工作目前仍处在初步应用阶段，未来仍有较大探索提升空间，针对现在存在的一些短板问题，建议可从如下几个方面进行优化：

(1) 完善舆情分析及监控模型，打造场景化舆情分析的应用闭环。例如行业监管领域突破“被动监管”，建立“异常对象”舆情监控机制，从而实现舆情监控的事中监管。

(2) 形成内外部数据融合分析，提高业务处理精准性。优化内外部数据的关联验证手段，提升结论的业务可解释性。在对外部舆情信息分析基础上，增加更多维度的内部数据验证，增强结果的业务可信度。

(3) 优化舆情处理基础技术架构，提升舆情数据的实时分析能力。可利用大数据平台提高数据处理能力、不断优化性能，探索应用新的技术框架提升舆情处理能力、提高舆情分析的时效性。

(责任编辑：赵博)