Computer Science and Application 计算机科学与应用, 2025, 15(10), 372-378
Published Online October 2025 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2025.1510276

Hans汉斯

基于BERT与DeepSeek大模型的智能舆论监控系统设计

韦玉荣*，刘小满#，熊圣昊，吴　桢

广西民族师范学院数学与计算机科学学院，广西　崇左

收稿日期：2025年9月22日；录用日期：2025年10月23日；发布日期：2025年10月31日

摘　要

本研究基于BERT模型与DeepSeek大模型，构建了一个智能舆情监测系统。该系统总体架构分为数据采集层、情感分析层、可视化交互层和智能报告层，技术实现上融合了微调BERT模型、Tkinter图形界面以及多源API集成。数据流程涵盖从光明网、Coze等多平台舆情信息采集，到基于公司金融领域微调BERT模型的情感自动标注，再到多维度数据可视化与DeepSeek生成的智能舆情分析报告。系统功能集成舆情动态抓取、情感分类、可视化展示与报告生成四大模块，实现了从数据获取到决策建议的全流程自动化。该系统的建设为舆情监控与风险应对提供了基于深度学习的智能支持，有助于提升企业对突发舆情的响应速度与决策科学性。

关键词

BERT，Coze新闻爬取，DeepSeek大模型，Tkinter，数据可视化

Design of an Intelligent Public Opinion Monitoring System Based on the BERT and DeepSeek Large Model

Yurong Wei*, Xiaoman Liu#, Shenghao Xiong, Zhen Wu

School of Mathematics and Computer Science, Guangxi Minzu Normal University, Chongzuo Guangxi

Received: September 22, 2025; accepted: October 23, 2025; published: October 31, 2025

*第一作者。
#通讯作者。

文章引用: 韦玉荣, 刘小满, 熊圣昊, 吴桢. 基于 BERT 与 DeepSeek 大模型的智能舆论监控系统设计[J]. 计算机科学与应用, 2025, 15(10): 372-378. DOI: 10.12677/csa.2025.1510276

Abstract

This study constructed an intelligent public opinion monitoring system based on the BERT and DeepSeek large models. The system's overall architecture consists of a data collection layer, a sentiment analysis layer, a visualization and interaction layer, and an intelligent reporting layer. Its technical implementation integrates a fine-tuned BERT model, a Tkinter graphical interface, and multi-source API integration. The data pipeline encompasses the collection of public opinion information from multiple platforms, including Guangming.com and Coze, automatic sentiment annotation based on a fine-tuned BERT model for corporate finance, and multi-dimensional data visualization and intelligent public opinion analysis reports generated using DeepSeek. The system integrates four functional modules: dynamic public opinion capture, sentiment classification, visualization, and report generation, automating the entire process from data acquisition to decision-making recommendations. This system provides deep learning-based intelligent support for public opinion monitoring and risk response, helping to improve enterprises' response speed to sudden public opinion incidents and the scientific nature of their decision-making.

Keywords

BERT, Coze News Crawling, DeepSeek Large Model, Tkinter, Data Visualization

1. 引言

随着互联网与社交媒体平台的深度普及，网络舆情已成为反映社情民意、影响社会稳定与企业发展的重要风向标。据中国互联网络信息中心统计，2024 年网络舆情日均增量高达 12.7 万条[1]，信息规模的爆炸式增长对舆情监测的实时性、准确性与智能化水平提出了前所未有的挑战。在此背景下，智能舆论监控系统已成为共同关注的研究热点。

在国际上，基于深度学习的自然语言处理技术已成为舆情分析的主流范式。特别是以 BERT 为代表的预训练模型，凭借其强大的语义表征能力，在情感分析等任务中展现出显著优势。根据 ACL 2024 会议报告，BERT 等模型在情感分析中的准确率已显著超越传统 LSTM 模型。发达国家的研究机构与商业公司(如 Brandwatch、Cision)已率先将大模型技术深度集成到其商业化舆情产品中。相比之下，国内智能舆情监控研究与应用虽发展迅速，但仍面临若干瓶颈。首先，通用预训练模型(如通用领域 BERT)在特定垂直领域(如金融、法律)的专业术语和语义理解上存在不足。同时现有解决方案成本高昂，商业舆情系统(如清博、鹰眼)的 API 调用成本通常高达 3~5 元/次，这严重阻碍了广大中小企业对先进舆情分析工具的应用。因此我们运用当下最热门的 DeepSeek 大模型与 Coze 平台等技术搭建了一个新闻舆情监控系统，实现了全流程的新闻舆情监控。

2. 数据采集模块

2.1. 双平台爬取新闻

爬取的网站选择的是光明网。传统人工获取数据的方式效率低下，难以满足大规模数据采集的需求。因此我们采用基于 Selenium 的自动化数据爬取技术。Selenium 是一种浏览器自动化测试工具[2]，可以模

拟真实用户操作浏览器的过程。相比传统人工或静态爬虫，Selenium [3]显著提升了数据获取的效率和覆盖率。最后返回的结果是 content (新闻内容)，source (来源)，time (时间)三方面的信息。

在 Coze 中设置爬取各方网址的新闻工作流如上图 1 所示。Coze API 能够实时抓取各大新闻平台的最新数[4]，具有极强的时效性优势。通过智能化的数据采集和处理机制，Coze 可以确保获取的舆情信息始终与各平台保持同步更新。

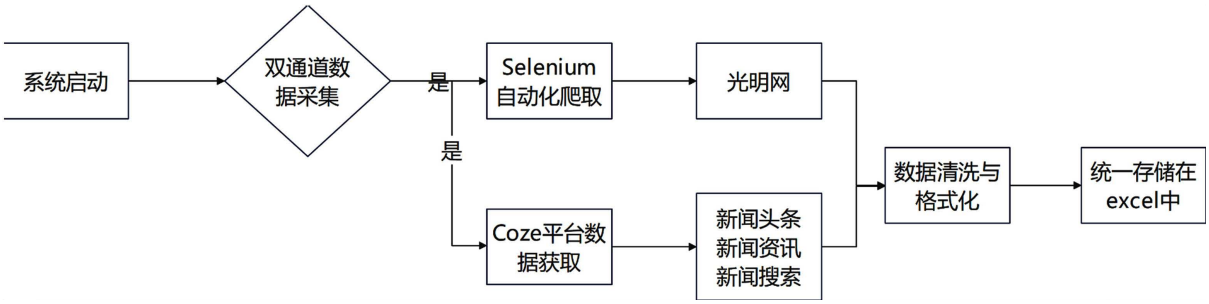

Figure 1. Dual-pipeline processing flow chart
图 1. 双管道处理流程图

2.2. 可视化获取数据

系统通过两个渠道进行采集舆情数据，将光明网和 Coze 平台获取的新闻内容、来源及时间等关键信息统一整合至 DataFrame 数据结构中，并最终导出为 Excel 文件(如表 1)。均能稳定获取舆情监测所需的核心字段(内容、来源、时间)，完全满足目前的舆情监控工作的基础需求。爬取的新闻数据可以自主地选择下载的路径。

Table 1. Results of automatic dual-channel information collection
表 1. 双渠道信息自动化采集结果

content	source	time
“京标”扩容北京车市抢跑发放窗口期小米汽车销售人员龚军(化名)便在微信朋友圈发布相应信息……	北京商报	2025-05-27 12:49:36
小米发布首款 3 nm 芯片累计投入已超 135 亿元当天，小米汽车首款 SUV 车型小米 YU7 首次正式亮相……	光明网	2025-05-22 21:54:16
……		
2025 中国汽车工程学会年会前瞻：一系列“首次”“首发”将亮相目前确定参加展览的重点整车企业有：比业迪、赛力斯、长安汽车、东风汽车、吉利汽车、小鹏汽车、小米汽车、零跑汽车等……	光明网	2025-05-21 10:03:02

3. 情感分析模块

3.1. 模型训练数据

该数据集来源于飞桨 AI Studio 平台的公开情感分析语料库，已由平台进行了初步的情感标注。尽管源数据集已包含基础的情感标注，为确保标注质量的可靠性及其与本研究任务的一致性，执行系统的质量控制。采用明确的三分类情感分析体系对文本进行界定：积极指文本中包含对上市公司股价、业绩、管理层、产品或未来发展前景的看好、赞扬、乐观期待等情绪；中性指文本仅为客观的事实陈述、新闻

转载或询问，不带有明显的主观情感色彩；消极指文本中包含对上市公司的批评、质疑、担忧、失望或看空等负面情绪。

质量控制过程包含四个关键步骤：首先，采用随机抽样方法从原始数据中抽取 10%的样本(约 3000 条)进行验证；其次，由两名研究人员遵循统一的标注标准独立进行重新标注，实现交叉检验；针对标注不一致的样本，引入第三名研究人员担任仲裁角色进行最终裁定；根据验证结果对发现的系统性标注错误进行修正。通过这一严谨流程，显著提升了数据集的标注质量与一致性，为后续模型训练奠定了可靠基础。经过质量检验后的最终数据集统计特征如下表 2 所示。

Table 2. Dataset category distribution statistics
表 2. 数据集类别分布统计

总样本数	31,226
积极	16,857 条(54.0%)
中性	7523 条(24.1%)
消极	6846 条(21.9%)

原始社交媒体文本包含大量噪声，必须经过严格的清洗和标准化流程才能用于模型训练。我们采用了以下多步骤预处理：

step1：格式标准化：使用正则表达式将所有非标准表情符号(如“[笑 cry]”)统一转换为中文括号格式(如“【笑 cry】”)，将其视为一个特殊的情感词汇予以保留。

Step2：无关信息剔除：彻底移除所有 URL 链接、HTML 标签以及`@用户`和`//@`(转发)信息，这些信息对情感判断没有贡献。

Step3：特殊字符处理：清理非中英文文字、数字和基本标点符号(“!?，。：【】”)以外的所有特殊字符，并将连续重复的标点符号压缩为单个。

Step4：停用词过滤：加载了包含 1893 个词条的中文停用词表(包括常见虚词、语气助词等)，并在分词后将其从文本中剔除，以降低特征空间的维度并突出关键情感词。

经过预处理后，我们得到了一个包含约 3.1 万条的评论文本的数据集。所有样本被按 72%:8%:20%的比例随机划分为。

3.2. BERT 模型微调训练

3.2.1. 预训练模型选择与架构

本研究采用“bert-base-Chinese”版本作为基础预训练模型[5] (由 Google Research 发布)。该模型是一个基于 Transformer 架构的双向编码表示模型，专为处理中文文本设计。其核心参数配置如下：12 层 Transformer 层、768 维隐藏层、12 个注意力头，共计约 1.02 亿参数。选择该模型的原因在于其已在大规模通用中文语料上进行了充分的预训练，捕获了丰富的中文语法和语义表征，为下游金融情感分析任务的领域自适应提供了强大的先验知识基础。

3.2.2. 领域适配与数据不平衡处理

为将通用预训练模型适配于金融舆情领域，于是对原始文本数据进行了预处理见 3.1 模块。处理完成之后进行数据的标签映射将中文情感标签(“积极”、“中性”、“消极”)映射为数字标签(0, 1, 2)。针对金融舆情数据中常见的类别不平衡问题[6] (本数据集分布约为 54%:24%:22%)，采用了代价敏感学习策略，自动计算类别权重，在训练过程中赋予少数类别更高的损失权重，从而迫使模型更加关注难以分类

的样本，缓解模型偏见。

$$w_j = \frac{N}{c \times N_j}$$

其中，w_j 表示类别 j 的权重，N 为训练样本总数，c 为类别总数，N_j 为类别 j 的样本数量。计算出各个比例的样本权重：积极权重为 0.617，中性权重为 1.384，消极权重为 1.520。

3.2.3. 微调超参数配置

为确保预训练 BERT 模型在金融舆情情感分析任务上的最优性能，本研究通过多轮实验确定了关键超参数的配置。模型微调采用小批量梯度下降策略，批处理大小(Batch Size)设置为 20，在计算效率与梯度稳定性之间取得平衡。训练周期(Epochs)设定为 2 轮，以避免过拟合并确保模型充分收敛。优化器选用 AdamW 算法，其学习率设置为 2×10^{-5}，该值是 BERT 模型微调的典型取值，既能有效更新权重又不破坏预训练获得的语言先验知识。文本序列最大长度(Max Length)截断为 40 个字符，基于金融新闻标题与短评论的语料特征分析而定。针对训练数据中存在的类别不平衡问题(积极 54.0%、中性 24.1%、消极 21.9%)，采用代价敏感学习机制，自动计算类别权重(积极：0.617，中性：1.384，消极：1.520)，在损失函数中赋予少数类别更高权重，以提升模型对负面情感的识别灵敏度。所有实验均设置随机种子为 42，确保结果的可复现性。具体流程见图 2。

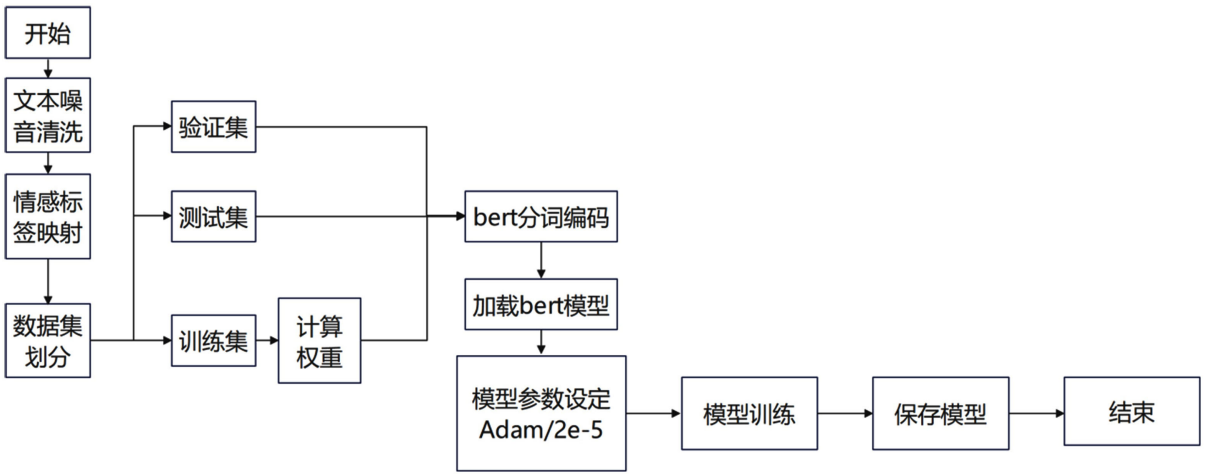

Figure 2. Model fine-tuning flow chart
图 2. 模型微调流程图

3.2.4. 微调流程与评估

最终测试集上的 79.76%准确率与验证集性能高度一致，证明了模型的良好泛化能力。在金融舆情监控的实际应用中，这一准确率水平已经具备实用价值，能够为投资者和监管机构提供可靠的情感倾向判断。值得注意的是，测试损失为 0.5324，与验证损失 0.5524 接近，进一步确认了模型在未知数据上的稳定性。

4. 舆情分析及报告下载功能

4.1. 可视化结果分析

在可视化分析中我们采用了基本的饼图和堆积柱形图，进行数据的统计计算占比以及展现排名。词云图能够直观地展示新闻报道中的高频词汇，帮助读者快速捕捉到当前热点话题或事件的核心内容。其

次，词云图可以揭示新闻报道的情感倾向或主题分布。最后，词云的美学设计能够增强新闻的可读性和吸引力，吸引更多读者关注。因此，词云图不仅是新闻数据分析的有力工具，也是提升信息传播效果的重要手段。

4.2. 舆论报告

舆情分析报告导出功能通过自动化技术将复杂的舆论数据转化为结构化报告，帮助用户快速掌握舆论趋势。通过自定义分析需求、自动化数据处理和 AI 智能生成，系统实现了高效、精准的舆情分析。报告采用标准化 Markdown 格式和时间戳命名，确保内容清晰且可追溯。

5. 系统页面设计

系统页面设计如下图 3。

系统界面设计

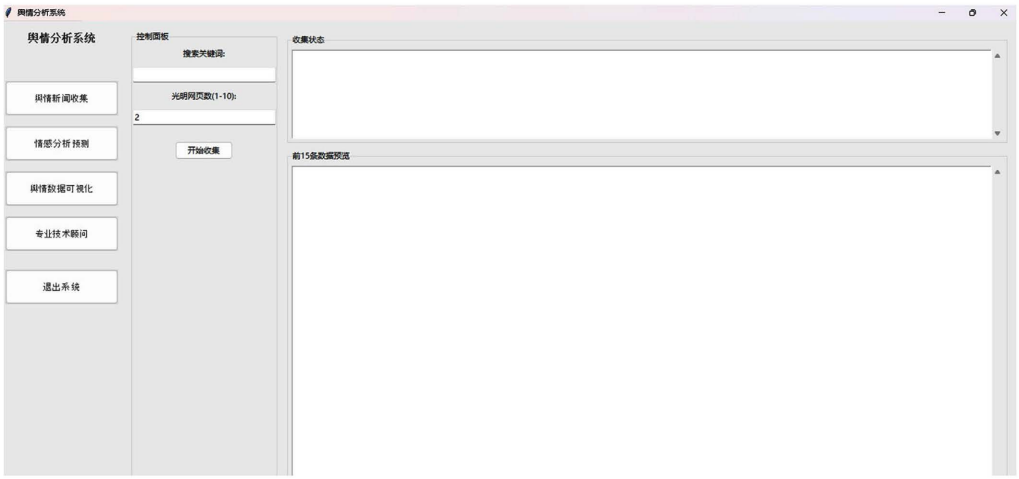

Figure 3. System page design
图 3. 系统页面设计

6. 结语

首先，项目开发了动态交互式可视化系统，通过情感卡片流、实时词云还有多维统计图表，直观查看到了舆情分布态势，大幅提升了数据可解释性；其次，针对舆情特点对 BERT 模型进行公司领域的微调，引入公司领域语料和优化损失函数，提升情感分类准确率；最后，创造性接入 DeepSeek 大模型还有 Coze 平台，构建了 AI 驱动的智能报告生成模块，能够自动产出包含趋势分析、风险预警和应对建议的结构化报告，将传统人工分析 8 小时的工作量压缩至 3 分钟内完成。这三个创新点形成“数据可视化－智能分析－决策支持”的完整技术闭环，既体现了对深度学习前沿技术的应用，又充分考虑了实际业务场景的需求，为舆情监测领域提供了可落地的智能化解决方案。

基金项目

基金课题：2025 年广西壮族自治区大学生创新创业训练计划《基于 BERT 模型与 DeepSeek 模型的舆论监控系统》(编号：S202510604089)阶段性成果。

参考文献

[1] 张昕, 范广宇. 主流媒体开展舆论监督报道的思考[J]. 记者摇篮, 2025(3): 27-29.

[2] 马孝宗. 基于 Python 实现数采系统实时数据的定时采集与处理[J]. 电脑编程技巧与维护, 2019(11): 42-45.

[3] 杜彬. 基于 Selenium 的定向网络爬虫设计与实现[J]. 金融科技时代, 2016(7): 35-39.

[4] 何科, 江雅珍, 李良晨. 基于 Coze 平台构建的锻造仿真软件的智能问答工作流研究——大语言模型与结构化知识库的协同应用[J]. 锻造与冲压, 2025(7): 20+22+24.

[5] 林原, 张亚, 于蒙, 许侃, 林鸿飞. 基于预训练模型的仇恨言论检测[J/OL]. 山东大学学报(理学版), 1-10. https://link.cnki.net/urlid/37.1389.N.20250408.1648.002, 2025-04-09.

[6] 陆钦华, 陈嘉宇, 王旭航, 等. 数据不平衡故障诊断: 一种预训练数据增强方法[J]. 测控技术, 2025, 44(1): 10-21.