韩磊, 夏明亮, 施展, 等. 基于大数据技术的高校舆情分析模型研究[J]. 智能计算机与应用,2024,14(11):194-199. DOI: 10.20169/j.issn.2095-2163.241130

# 基于大数据技术的高校舆情分析模型研究

韩 磊, 夏明亮, 施 展, 郑胜男
(南京工程学院 计算机工程学院, 南京 211167)

**摘 要**: 舆情分析系统的研发是辅助高校舆情治理的重要方式。针对现有系统在技术架构、数据采集和分析方面的不足,设计了基于流数据的舆情采集和存储技术框架,实现了基于 LDA(Latent Dirichlet Allocation)的热点主题挖掘方法,提出了基于 BERT(Bidirectional Encoder Representations from Transformers)预训练模型的情感分类方法,搭建了面向高校舆情分析的 Web 系统。为高校舆情分析系统的设计和实现提供有益的参考和解决方案。

**关键词**: 舆情治理; 大数据; 情感分类; 主题挖掘
**中图分类号**: TP391 **文献标志码**: A **文章编号**: 2095-2163(2024)11-0194-06

## Research on public opinion analysis model based on big data technology

HAN Lei, XIA Mingliang, SHI Zhan, ZHENG Shengnan
(School of Computer Science Engineering, Nanjing Institute of Technology, Nanjing 211167, China)

**Abstract**: The public opinion analysis system plays a crucial role in assisting university public opinion governance. However, existing systems often face limitations in their technical architecture, data collection, and analysis capabilities. To address these shortcomings, the paper has developed a comprehensive framework for sentiment analysis and data storage based on streaming data. The paper has also implemented a cutting-edge method for mining hot topics using Latent Dirichlet Allocation (LDA) and introduces a sentiment classification approach leveraging the power of the BERT (Bidirectional Encoder Representations from Transformers) pre-trained model. Furthermore, the paper has successfully constructed a user-friendly Web-based system tailored specifically for university public opinion analysis. The research provides valuable insights, innovative solutions, and practical recommendations for the design and implementation of effective university public opinion analysis systems.

**Key words**: public opinion governance; big data; sentiment classification; topic mining

## 0 引 言

随着互联网和社交媒体的快速发展,高校的舆情工作也面临着前所未有的挑战和机遇。对于高校来说,舆情的有效分析和管理对于提升学校声誉、解决潜在危机、塑造学校形象具有重要意义[1]。然而,由于舆情信息的多样性和庞杂性,传统的手工分析方法已经无法满足高校舆情管理的需求,而当前使用的舆情分析系统也存在一些不足之处。

首先,现有的舆情分析系统在数据采集方面存在局限性。虽然社交媒体平台和网络论坛等渠道提供了大量的舆情数据,但数据的获取、整合和存储仍有不少难题待解,导致分析结果的完整性和准确性受到限制[2]。

其次,现有系统在舆情分析的深度和广度上存在局限性。舆情分析需要从海量的文本数据中提取有价值的信息,包括情感倾向、关键主题等。然而,现有系统在情感分析、文本挖掘和主题识别等方面的算法和模型相对简单,无法充分挖掘数据中的潜在信息,导致后期对舆情的深入理解和分析受到限制。

为此,本文基于大数据技术设计高校舆情分析模型,以辅助高校有效地监测、分析和应对舆情事件。该系统模型利用大数据技术,结合文本挖掘、情感分析和机器学习算法,从社交媒体平台、新闻网站和论坛等渠道采集、整理海量的舆情数据。通过对舆情数据的深度挖掘和分析,系统模型能够实时监

**基金项目**: 江苏省高校哲学社会科学研究项目(2022SJYB0436); 江苏省智能感知技术与装备工程研究中心开放基金项目(ITS202101, ITS202201)。
**作者简介**: 韩 磊(1982—),男,博士,副教授,主要研究方向:大数据技术,计算机视觉。Email:hanl@njit.edu.cn; 夏明亮(1969—),男,高级工程师,主要研究方向:大数据技术,智能制造; 施 展(1983—),男,博士,副教授,主要研究方向:人工智能,大数据技术; 郑胜男(1986—),女,实验师,主要研究方向:人工智能,大数据分析。
**收稿日期**: 2023-06-20

测高校舆情动态，提供全面的舆情评估和情感分析报告，为高校提供科学的决策依据和支持。

## 1　相关理论及方法

### 1.1　文本处理技术

本文舆情数据主要是文本数据，文本处理技术在高校舆情分析系统中起着重要作用，主要涉及如下技术。

（1）文本预处理：文本预处理是文本处理的第一步，旨在清洗和规范原始文本数据。主要包括去除标点符号、停用词、数字和特殊字符，进行大小写转换，以及词干提取和词形还原等操作[3]。文本预处理有助于减少数据噪音，简化文本表示，提高后续文本处理任务的效果。

（2）词袋模型（Bag-of-Words，BoW）：词袋模型将文本表示为词语的集合，忽略了词语的顺序和语法信息，只关注词语出现的频率。词袋模型能够有效地捕捉文本中的关键词信息，用于后续的情感分析、主题识别等任务[4]。

### 1.2　注意力机制

注意力机制（Attention Mechanism）是一种在深度学习中常用的技术，用于加强模型对输入数据的关注程度[5]。通过模拟人类的注意力机制，使得模型能够将注意力动态地集中在输入的不同部分上。

在自然语言处理中，注意力机制在机器翻译、文本摘要和情感分析等方面获得了广泛应用。其中，BERT（Bidirectional Encoder Representations from Transformers）模型是一种基于注意力机制的预训练模型，在情感分析任务中凭借其性能特点，展现出了显著优势[6]。

BERT 模型通过预训练的方式，在大规模的语料库上学习语言模型。能够利用双向 Transformer 模型来编码输入文本的上下文信息，从而深度理解文本中的语义和句法关系。BERT 模型还引入了多层自注意力机制，使得模型在编码文本时能够根据输入序列的不同部分自适应地分配注意力权重。

在情感分析任务中，BERT 模型可以通过微调（fine-tuning）来适应特定的情感分类任务。通过将情感分类任务的标签信息与 BERT 模型结合，可以将 BERT 模型转化为一个情感分析模型。由于 BERT 模型在预训练阶段已经学习到了丰富的语言表示，微调后的模型能够更准确地捕捉文本中的情感信息，提高情感分析的性能。

### 1.3　模型评价指标

情感分类模型是舆情分析的关键[7-8]，评价情感分类模型的指标通常有如下 3 种。

（1）精确率（*Precision*）：精确率衡量的是模型预测为正例的样本中，真正为正例的比例。精确率计算公式为：

$$Precision = TP/(TP + FP) \tag{1}$$

其中，*TP* 表示真正例，*FP* 表示假正例。

精确率关注的是模型的正例预测能力，适用于对假正例有较高敏感性的情况。

（2）召回率（*Recall*）：召回率衡量的是模型对正例样本的识别能力，即模型能够正确预测出多少正例样本。召回率计算公式为：

$$Recall = TP/(TP + FN) \tag{2}$$

其中，*TP* 表示真正例，*FN* 表示假反例。

召回率适用于对假反例有较高敏感性的情况。

（3）*F*1 值（*F*1 - Score）：*F*1 值是精确率和召回率的调和平均数，综合考虑了两者的性能。*F*1 值能够综合衡量模型的准确率和召回率，并在两者之间取得一个平衡。

## 2　高校舆情分析模型设计

### 2.1　框架设计

本文基于大数据技术实现舆情采集、分析和展示，系统功能与 Web 服务框架融为一体。舆情分析功能模型架构如图 1 所示。由图 1 看到用户通过前端访问系统，可进行系统配置和用户管理，并通过前端交互实现舆情管理、分析及舆情报告生成等。业务逻辑层提供舆情分析的服务支持，包括网页爬取、数据清洗、数据存储、主题分析和情感分析等。数据存储包括数据访问层和数据库存储，数据访问层又称持久化层，主要负责对数据的访问；考虑到网络爬虫所采集的多为文档数据，而数据清洗和分析后又可形成结构化数据，所以本文系统中使用了 MongoDB、MySQL 和 Redis 三种数据库。

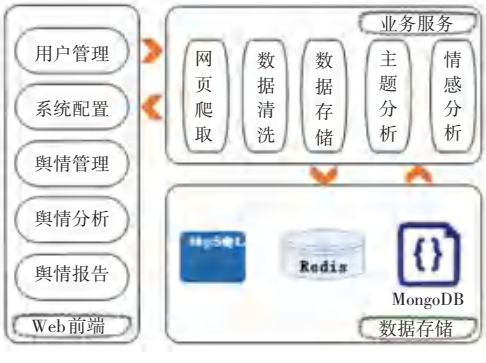


**图 1　舆情分析功能模型**

**Fig. 1　Public opinion analysis functional model**

根据上述功能模型,结合大数据平台技术,本文的技术选型如图 2 所示。由图 2 可知,采用 Scrapy 爬取百度贴吧、新浪微博等舆情信息,然后采用 Kafka 接收流式数据;再发送至 Spark Streaming 进行微批量流处理,依托 Spark 大平台,使用 Spark Streaming 既统一了技术堆栈,又确保了与其他 Spark 组件的无缝交互,实现了与 MongoDB、Redis 和 MySQL 的交互。在此基础上,搭建 TensorFlowOnSpark 实现舆情主题分析和情感倾向分析。最后,以 Web 前端应用的形式实现数据可视化与应用交互。

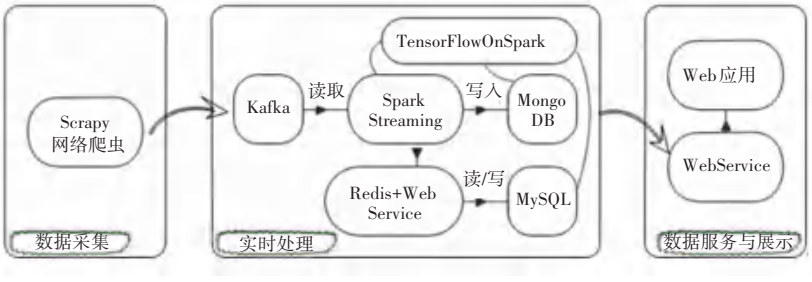


**图 2 舆情分析技术框架**

**Fig. 2 Technical framework for public opinion analysis**

## 2.2 模块设计

### 2.2.1 舆情数据采集与清洗

舆情数据的自动采集是舆情分析系统的基础模块。为提高系统的适应性,系统留有数据源的配置界面,支持百度贴吧、新浪微博和主流新闻网站的选择。根据用户的配置,构建 URL 地址库。

网页信息采集如图 3 所示。图 3 中,爬虫程序选取地址库中的地址,根据相关配置和既定访问方式爬取顶层网页,并建立快照库;然后对快照库中内容进行信息提取或深层 URL 地址解析;提取到的信息送入流处理模块,深层 URL 地址送入地址库,在下一轮迭代中进行深层网页爬取。以百度贴吧为例,爬虫自动遍历帖子、回复和评论等内容,获取用户发表的言论和讨论。

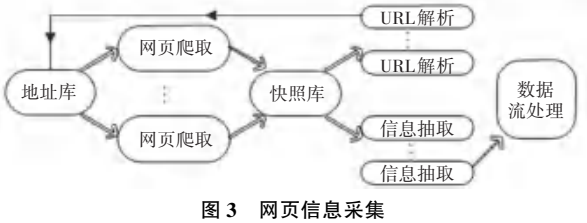


**图 3 网页信息采集**

**Fig. 3 Web page information collection**

对爬取的数据需要进行筛选和过滤,以提取与高校舆情相关的信息。这可以通过关键词匹配、主题分类等方法来实现。只有与高校相关的帖子和评论会被保留,其他非相关信息将被滤除。

在数据清洗阶段,需要处理数据中的噪声和错误,以确保数据的质量和一致性。常见的数据清洗操作包括去除 HTML 标签、处理重复数据、纠正拼写错误和标点符号等。

### 2.2.2 高校舆情数据存储

舆情数据存储模块接收采集模块的数据,同时为模型训练、数据分析与展示提供支持。根据高校舆情分析系统的数据需求分析,系统中的数据主要有模型训练数据、帖子和评论数据、元数据、分析过程或结果数据等,本文选用 MySQL、Redis 和 MongoDB 构建存储系统,各数据库在系统中的职责和关系如下。

(1) MySQL。MySQL 是一种关系型数据库,适合存储结构化的数据。在高校舆情分析系统中,MySQL 可用于存储以下类型的数据:

① 用户信息:存储高校舆情系统的用户信息,包括用户名、密码、角色等。

② 帖子的元数据:存储帖子的元数据,如帖子的 ID、标题、发布时间、作者 ID 等。

MySQL 的表格结构和关系型数据库的特性使其适合存储和管理高校舆情数据中的结构化信息。MySQL 提供了复杂的查询和事务处理功能,可以进行复杂关系和关联查询,以支持系统中的数据分析和用户需求。

(2) MongoDB。MongoDB 是一种文档型数据库,适合存储非结构化或半结构化的数据。在高校舆情分析系统中,MongoDB 可用于存储以下类型的数据:

① 帖子内容:每个帖子可以表示为一个文档,其中包含帖子的标题、内容、发布时间、作者等信息。

② 评论数据:每条评论也可以表示为一个文档,其中包含评论的内容、发布时间、作者等信息。

MongoDB 的灵活的文档模型和强大的查询能

力使其适合存储和检索高校舆情数据中的帖子和评论等非结构化信息。此外，MongoDB 还支持分布式架构，可以进行数据的水平扩展，以满足大规模数据存储的需求。

（3）Redis。Redis 是一种内存型键值数据库，适合高速读写和缓存操作。在高校舆情分析系统中，Redis 可用于存储以下类型的数据：

① 临时数据：存储临时性的数据，如临时计算结果、中间状态等。

② 统计结果：存储高校舆情数据的统计结果，如某个主题的热度、关键词的频率等。

③ 训练数据：为提高训练效率，将用于训练文本情感分类的数据集加载到 Redis 数据库。

Redis 的主要特点是快速读写和低延迟，适用于实时数据的存储和查询需求。通过将临时数据和统计结果存储在 Redis 中，系统可以快速访问这些数据，提高系统的性能和响应速度。

高校舆情分析系统的数据存储需求涵盖了非结构化或半结构化的帖子和评论数据、结构化的用户信息和帖子元数据、临时数据和统计结果等。通过合理选择适用的数据库（如 MongoDB、MySQL 和 Redis）来存储不同类型的数据，可以满足系统对数据的存储、查询和分析的需求，并提供高性能和可扩展性的支持。

### 2.2.3　热点主题挖掘

对采集舆情数据进行主题挖掘，以利于进行热点追踪和情感分类。本文的热点主题挖掘流程如图 4 所示。

（1）需要对高校舆情数据进行预处理，包括去除停用词、进行分词、词干化等操作。这是为了将文本数据转化为可供主题建模算法处理的格式。

（2）构建文本集合。将预处理后的文本数据组成一个文本集合，每个文本代表一个帖子、评论或新闻等。这个文本集合将作为主题建模算法的输入。

（3）将文本集合送入 Latent Dirichlet Allocation（LDA）模型，进行分配与关联[9]。利用模型推断每个文本所属的主题分布。通过分析舆情数据中不同主题的分布，可以识别和挖掘出当前的热点主题。

Latent Dirichlet Allocation（LDA）是一种常用的主题建模方法，用于发现文本数据中的隐藏主题结构。LDA 的基本思想是，假设每个文档都是由多个主题以一定的概率分布生成的，而每个主题又是由一组词语以一定的概率分布生成的。具体步骤如下。

（1）初始化：为每个文档中的每个词随机指定一个主题。

（2）迭代训练：通过迭代的方式学习主题和词语的分布。在每一次迭代中，根据当前的主题分布和词语分布，计算每个词语属于每个主题的概率。然后根据这些概率重新分配每个词语的主题，直到收敛为止。

（3）输出结果：在训练结束后，可以得到每个文档的主题分布以及每个主题的词语分布，从而理解文本数据中的主题结构。

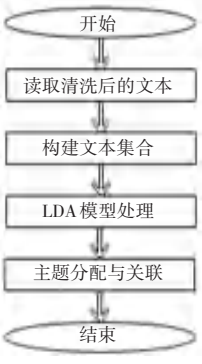


**图 4　主题挖掘流程**

**Fig. 4　Flowchart of topic mining**

### 2.2.4　舆情情感分类

舆情情感分类是对舆情事件做出判断，实时预警的关键[10]。本文基于 BERT 预训练模型设计舆情情感分类神经网络，其结构如图 5 所示。

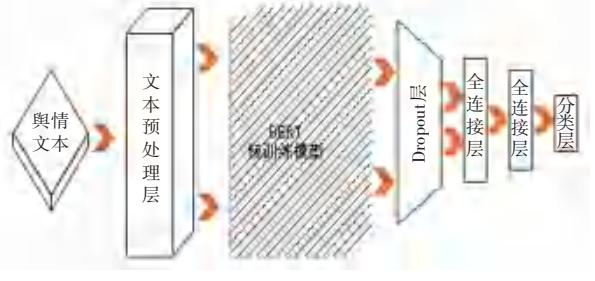


**图 5　舆情情感分类神经网络**

**Fig. 5　Neural network for classification of public opinion**

输入数据通过 BERT 预训练模型进行编码，经过 Dropout 层和全连接层后，通过 Softmax 函数进行概率计算，最后输出情感分类结果。该网络的结构包括以下组件。

（1）文本预处理层：主要使清洗后的舆情文本与 BERT 预训练模型适配。

（2）BERT 预训练模型：用于从输入文本中提取特征。

（3）Dropout 层：用于防止过拟合，随机丢弃一

部分神经元。

(4)全连接层:用于将 BERT 的输出特征映射到情感分类的标签空间。

(5)分类层:运用 Softmax 函数将输出的连续值转换为概率分布,输出预测的情感分类结果。

在该网络中,BERT 预训练模型的输出张量格式是一个二维张量,具体表示为(*batch_size*, *hidden_size*)。其中,*batch_size* 为输入文本的批量大小,*hidden_size* 为 BERT 模型的隐藏状态的维度大小。在 BERT 模型中,每个输入句子会经过一系列的 Transformer 层,其中包含自注意力机制(self-attention)和前馈神经网络(feed-forward neural network)。这些层的作用是对输入句子进行特征提取和学习表示,将输入句子中的每个单词转化为其在语义空间中的表示。BERT 模型的最后一层是池化层,可将每个输入句子中的所有隐藏状态(每个单词对应一个隐藏状态)进行汇总,并生成一个全局的句子表示。这个句子表示在模型的输出中被称为“pooled_output”,对应于输出张量中的每个样本。

## 3 高校舆情分析系统实现与应用

### 3.1 功能实现

根据第 2 节的架构,本文分别以 Vue 和 SpringBoot 为前后端框架,基于 Java、Python、JavaScript、HTML、CSS 和 Shell 脚本实现系统编程,高校舆情分析系统界面如图 6 所示。


**图 6 高校舆情分析系统**

**Fig. 6 University public opinion analysis system**

用户通过高校舆情分析系统能实现舆情监测、舆情分析和舆情报告等功能交互,系统能够实时呈现热点话题、热点词云、情感倾向分析结果等。可以定期生成舆情报告,针对热点问题进行追踪发现等。

### 3.2 系统测试

分别在 online_shopping_10_cats、weibo_senti_100k 数据集上训练和测试本文的情感分类模型。其中,online_shopping_10_cats 数据集包含 10 个类别(书籍、平板、手机、水果、洗发水、热水器、蒙牛、衣服、计算机、酒店),共 6 万多条评论数据,正、负向评论各约 3 万条;Weibo_senti_100k 数据集的数据来源于新浪微博的各种评论,共 10 万多条数据。

选用精确率、召回率和 *F*1 分数三个指标评价本文的情感分类方法。其中,精确率是指在被所有预测为正的样本中实际为正样本的概率,召回率体现了实际为正的样本中被预测为正样本的概率,*F*1 分数则反映了两者的平衡点。测试结果见表 1,本文的情感分类模型在 2 个数据集上都取得了较好的效果,符合高校舆情分析系统情感分类的需求。

**表 1 情感分类算法测试**

**Table 1 Test results of sentiment classification algorithm** %

| 类别 | 精确率 | 召回率 | *F*1 分数 |
|---|---|---|---|
| online_shopping_10_cats | 93.02 | 93.56 | 93.28 |
| Weibo_senti_100k | 91.63 | 91.83 | 91.73 |

## 4　结束语

高校舆情分析系统的设计和实现具有重要意义。通过高校舆情分析系统，可以深入挖掘和分析高校舆情数据，帮助高校更好地了解学生、教职员工和社会公众的关切问题和当下需求。本文采用 Web 前端、业务服务和数据存储 3 层功能模型，利用 MongoDB、MySQL 和 Redis 等数据库的组合，以及 Spark Streaming 进行流数据处理，搭建了一个高效可靠的数据存储和处理系统，实现对大规模高校舆情数据的快速存取和分析。在 TensorFlowOnSpark 基础上，引入文本处理技术和情感分类方法，实现高校舆情数据的精细化分析，深入挖掘其中的情感倾向和热点主题。本文系统实现了对高校舆情数据的深度剖释及洞察，这将为高校提供更全面的舆情分析和决策支持，促进高校舆情管理的科学化和精细化发展。

## 参考文献

[1] 赵好好. 基于数据挖掘的高校网络舆情热点话题分析及引导策略研究[J]. 信息系统工程, 2023(2): 117-119.

[2] ZHU Yalin, LI Xiangwei, WANG Juan. Analysis and research of weibo public opinion based on text [J]. Journal of Physics Conference Series, 2021, 1769(1):012018.

[3] 周春梅，冯林，张华辉. 网络舆情对新冠疫情下青少年情感态度的分析[J]. 计算机仿真, 2023, 40(1): 553-558.

[4] 程名，刘虎，郑涵予，等. 舆情大数据在 PMI 数据分析预测中的应用研究[J]. 中国国情国力, 2023(2): 44-50.

[5] 曾莉，杨添宝，周慧. 基于 LDA 与注意力机制 BiLSTM 的微博舆情分析模型[J]. 南京理工大学学报, 2022, 46(6): 742-748.

[6] 满瑞. 基于 CNN-BILSTM 网络与 BERT 的舆情分析算法[D]. 桂林：桂林电子科技大学，2021.

[7] ZHOU Zhipeng, ZHOU Xingnan, QIAN Lingfei. Online public opinion analysis on infrastructure megaprojects: toward an analytical framework [J]. Journal of Management in Engineering, 2021, 37 (1) : 04020105.

[8] 任伟建，刘圆圆，计妍，等. 基于 RNN-LSTM 新冠肺炎疫情下的微博舆情分析[J]. 吉林大学学报(信息科学版), 2022,40(4): 581-588.

[9] 钟金宏，黎梦萍，宣占祥. 基于 LDA 模型的高校疫情防控舆情分析[J]. 信息技术与信息化, 2022(12): 72-75.

[10] BARACHI M E, ALKHATIB M, MATHEW S, et al. A novel sentiment analysis framework for monitoring the evolving public opinion in real-time: Case study on climate change[J]. Journal of Cleaner Production, 2021,312: 127820.