**中国大学生计算机设计大赛**



大数据实践赛作品报告

作品编号：

作品名称：　　SIU: Sentiment Identification and Understanding

版本编号：

填写日期：　　　2023.3.26.

填写说明：

1. 本文档适用于大数据实践小类；
2. 正文一律用小四号宋体，1.3倍行距；一级标题为二号黑体，其他级别标题如有需要，可根据需要设置；
3. 本文档应结构清晰，突出重点，适当配合图表，描述准确，不易冗长拖沓；
4. 提交文档时，以PDF格式提交；
5. 本文档内容是正式参赛内容的组成部分，务必真实填写。如不属实，将导致奖项等级降低甚至终止本作品参加比赛。

**目 录**

[第1章 作品概述 1](#_Toc70706098)

[第2章 问题描述 1](#_Toc70706099)

[第3章 技术方案 1](#_Toc70706100)

[第4章 系统实现 1](#_Toc70706101)

[第5章 系统评测 1](#_Toc70706102)

[第6章 安装使用 2](#_Toc70706103)

[第7章 作品总结 2](#_Toc70706104)

[参考文献 2](#_Toc70706105)

# 作品概述

【填写说明：本部分非常重要，建议800字以内。简要说明作品的意义、技术特色、实现方法、运行（或应用）效果等等。着重介绍作品的特色、和运行（或应用）效果。】

## 1.1作品简介

我们使用预训练的BERT模型对中文文本进行情感分类，分析文本是正向的还是负向的。接着，使用爬虫从B站上爬取数据，使用此模型对文本进行分类。最后，对数据进行分析，例如对均值进行估计以判断舆情对某一方面的总体态度，对方差进行分析以判断舆情对某一个领域的态度的分化程度，以及对各个时间段上的舆情情况的变化进行分析等。

**1.2 作品意义**

方便快捷地对大量的舆情文本进行分析，可以帮助我们更好地了解人们对某个话题或事件的看法和反应。这对于政府、企业、媒体等各种机构来说都非常有用，可以帮助他们更好地了解公众的需求和意见，从而更好地制定政策、推出产品、进行宣传等。此外，这个作品也可以作为一个数据分析工具，帮助研究者更好地了解某个领域的舆情走向和热点问题，为后续研究提供有价值的参考和支持。

**1.3 实现方法**

1、情感分类模型：对具有预训练的二分类大语言模型**bert-base-chinese**进行微调，然后输出是0或1，代表其情感是正向还是负向；

2、爬虫：使用Python的Selenium库和Requests库，爬取B站各个tag的视频评论数据。其中，使用Selenium库用于模拟浏览器操作，使用Requests库用于发送网络请求获取数据。我们还实现了多个函数，主要功能包括获取各子模块视频链接、获取视频评论数据，以及将数据保存到本地文件中。

3、数据分析：使用Pandas进行统计分析，使用Matplotlib绘制图片进行可视化，使用Excel绘制更多数据可视化对图形，如饼图等。

**1.4 技术特色**

我们作品的技术特色非常明显，它不仅使用了BERT模型进行文本分析，还结合了爬虫技术，可以实时地从网络上爬取评论并对舆情进行分析。

BERT是一种预训练的自然语言处理模型，可以对大量的文本数据进行训练，从而学习到丰富的语言表达能力。使用BERT模型可以更好地理解和分析人们对某个话题或事件的看法和反应，从而为政府、企业、媒体等机构提供有用的数据支持。

同时，使用爬虫技术可以对指定的网站或社交媒体进行监测，实时地获取用户的评论和回复，并进行分析。通过结合BERT模型，可以对大量的舆情文本进行分析，从而更好地理解公众对某个话题或事件的看法和反应。

此外，我们的作品还使用了其他的统计方法和图表来细致地了解舆情变化和趋势。例如，热力图可以直观地展示不同地区或时间段的舆情热度；折线图和条形图则可以清晰地反映舆情的变化趋势和关键点。通过结合不同类型的图表，可以更加全面地了解舆情，为政策制定和决策提供有力的支持。

因此，我们的作品不仅仅是一个基于BERT模型的文本分析工具，而是一个综合性的数据分析工具，为政府、企业、媒体等机构提供了更好的了解公众需求和意见的途径，同时也为研究者提供了有价值的参考和支持。可以进一步拓展应用场景，例如在政策制定、市场营销、品牌塑造等方面发挥重要的作用。

**1.5 运行（或应用）效果**

# 问题描述

## 2.1 问题来源

【填写说明：说明问题的背景、起因等】

## 2.2 现有解决方案

【填写说明：分析现有类似的解决方案，或前人解决问题的途径（需标注参考引用），并进行分析；如果有同类竞品，建议从多个维度对本作品与竞品进行比较】

## 2.3 本作品要解决的痛点问题

【填写说明：基于2.2的对比分析，阐述本作品要解决的核心痛点问题】

## 2.4 解决问题的思路

【填写说明：作品的功能和性能需求；使用的数据集，包括数据格式，数据来源，数据获取方式，数据特点，数据规模等，并给出具体的数据样例。所提出的指标或要求必须在第5章得到印证】

# 技术方案

【填写说明：从原理层面，详细介绍系统所采用的技术方案，先总体介绍，给出技术路线框架图，然后分模块详细介绍。着重介绍解决问题的思路，以及所涉及的模型、算法等；原创工作详细描述，非原创工作简略描述，并尽可能标注引用文献】

# 系统实现

【填写说明：介绍系统的具体实现过程，特别是其中所遇到的困难，解决的方法等，这里只需要介绍团队真实发生的工作】

# 系统评测

【填写说明：通过测试与对比，论证系统的有效性，包括数据来源、数据规模、环境配置（包括但不限于操作系统、数据库、依赖的库文件、配置文件）、测试过程、分析与结论等等。各参赛队务必重视数据测试，所有对自己作品准确性、有效性、稳定性，甚至作品受欢迎的程度的宣称，都应该得到数据结果或对比实验的支持，确保作品的运行或使用具有可复现性，否则评审人有理由怀疑其真实性】

# 安装使用

【填写说明：简要说明安装环境要求、安装过程、主要操作方法程等，建议包含一些典型应用场景的使用说明。】

# 作品总结

【填写说明：从创意、技术路线、工作量、数据和测试效果等方面对作品进行自我评价和总结，并对作品的进一步提升和应用拓展提出展望】

# 参考文献

【请按照标准参考文件格式填写】