**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。

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 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】

互联网时代快速发展以来，网民数量不断扩大，根据第五十次《中国互联网发展状况统计报告》，截至2022年6月，我国网民规模为10.51亿，互联网普及率达74.4%。更由于互联网的发声门槛低的特性，人们常常更加愿意在网络上发表自己的意见而非传统渠道。因此，网络舆论已经变成社会舆论非常重要的一部分了。在此情况下，对网络舆论的分析就显得格外重要。

## 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】

目前大多数对情感倾向性分析的解决办法，都是首先对抓取的舆情数据进行分词处理，然后结合情感语料数据库和情感分析算法对切分后的语料进行情感计算、分析，并进行情感标注。最后通过聚类和分类得出个体情感倾向和群体情感倾向。这是一种从传统机器学习和计算语言学发展而来的分析方法。

## 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】

2.2中提到的现有的大多数情感倾向分析方法有几个问题：

1、对汉语的切割问题一直是一个老大难问题，传统机器学习方法效果并不是很好；

2、情感数据库的构建比较耗费人力，同时也比较麻烦；

3、由于1、的问题，通过聚类和分类对情感进行分析的方法效果在现在看来并不是很好。

## 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】

**作品功能和性能需求：**

1.模型的整个运作过程中，要尽量减少人力消耗，尽量做到一键式傻瓜操作

2.模型的情感分析效果要比较好。由于是大规模的数据分析，可以接受的平均准确率可以比人稍低，但是不能低太多。

**数据集：**

训练和测试数据集为互联网上开源的外卖评论数据集。总量有10000条。格式为.csv文件：

情感倾向（0或1）, 文本内容。

比如，有一条负面评价为：

0,多收了钱。

# 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】

所有有效的数据

部分人工

标注的数据

预训练BERT

Waimai\_10k.csv

输入数据

进行微调

微调后的BERT

爬虫

舆论数据

情感倾向数据

统计分析工具

舆情分析结论

小样本准确率

****一、BERT-base-chinese****

BERT (Bidirectional Encoder Representations from Transformers)是一种基于Transformers架构的预训练语言模型，在多个NLP任务上取得了最先进的性能。

BERT通过预先训练一个大型的无监督模型，使其能够理解自然语言中的上下文和语义信息。BERT模型包括两个阶段的训练：预训练和微调。在预训练阶段，BERT模型使用大量的未标记的文本数据来学习语言知识。在微调阶段，BERT模型使用标记的数据进行进一步的训练，以完成特定的任务。

bert-base-chinese是BERT的一个预训练模型，专门针对中文语言处理任务。它使用中文维基百科和百度百科作为预训练语料库进行训练，包含了21128个字形的汉字和21128个对应的字形的拼音。bert-base-chinese的预训练模型在中文NLP任务上表现出色，因此在中文自然语言处理领域中广受欢迎。

使用bert-base-chinese模型可以进行中文文本分类、序列标注、语言理解等多种NLP任务。BERT的出现和BERT-based模型的应用对于自然语言处理领域是一次重要的里程碑，因为它们在许多NLP任务上都取得了最先进的结果，为自然语言处理研究的未来提供了新的方向。

****2、爬虫****

****3、统计分析****

设是第个数据是否预测正确的变量，设是其正确率。即

1. 通过给模型喂一批小量的（数量为）数据样本，可以得到模型在这一批数据上的正确频率。使用点估计的做法，用此频率作为无偏估计其正确率，即。由于未知方差，则用样本方差来估计：
2. 在得到所有数据的分类情况以后，根据上文预估出来的准确率，我们可以得到一些近似的概率：

因此，如果设数据总量为，分类为正向情感的数量和负向情感的数量分别有,则实际上分类为正向的数量和负向的数量为

。并且，由于正向和负向的地位是等同的，两者的方差都应该是上文提到的。

1. 根据上文得到，我们可以分别得到正向和负向的舆论在这整个舆论场中所占比例。根据所占比例，我们可以分析出现在的舆论场是一边倒还是两边势均力敌。
2. 根据上文得到，我们可以用现实中发生的事件来对应各个时间段的舆论变化。
3. 我们可以再设一个随机变量。当第个数据的情感倾向是正向的时候，，反之。则有：
4. 显然，服从伯努利分布。设：

则对于，即。显然，服从二项分布，且其值代表了量化后的网络舆论对某事件的看法的情感倾向性。因此，我们有必要对的期望和方差进行区间估计。我们假设置信度水平。

1. 由于我们选择的比较大（），则根据中心极限定理的特例[[1]](#endnote-1)，我们可以用正态分布来近似表示这个二项分布，即。而正态分布的值可以直接查表，这使得计算方便。
2. 首先找一个期望的良好的点估计。我们将数据分为组，则选取。由于未知，不适合作枢轴变量，我们将样本分为类，选取样本标准差代替得到。则有:

改写为：

因此，其置信度水平为区间为：

1. 当其期望大于时，我们可以认为网民的态度偏正面；反之，则认为网民的态度偏负面。这是一个假设检验的过程，即：
   1. 有原假设和备择假设。现在给予抽样数据，以此来对此假设进行检验。
   2. 与区间估计同样的处理方法，则有

作为枢轴变量。

* 1. 假设显著性水平的情况下，我们得到检验：

# 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】

1. 陈希孺 概率论与数理统计[M].安徽: 中国科学技术大学出版社, 2009:145. [↑](#endnote-ref-1)