1. **当初为何选择这个选题？**

我们选择的是NLP文本情感分类，我们选择这个题目的原因之一是我们这个学期同时正在进行自然语言处理这门课，我认为这个选题可以更贴近我们的课程学习，同时，我们也可以实时把自然语言处理课程上的知识进行实际的运用，相佐相成。这个选题同时也是非常贴近我们的生活的，我们觉得让机器能够理解人类的语言情感是一件非常有美丽的事情，作为我们的第一个机器学习方面的项目，我们希望能够对这一方面有更多的了解，同时切实的体验。

**We chose NLP text sentiment classification because we are also taking a natural language processing course this semester. I believe this topic aligns well with our coursework, allowing us to apply what we're learning in a practical way. Additionally, this topic is very relevant to our daily lives. We think it’s fascinating to enable machines to understand human emotions through language. As our first machine learning project, we hope to gain deeper insights and hands-on experience in this field.**

1. **你们是如何分工的？**

A成员：核心代码实现

Core code development

职责：负责项目的代码编写，涵盖数据预处理、模型构建和训练等核心任务。

所需能力：具备扎实的Python编程技能，熟悉常用NLP工具（如NLTK、spaCy、Transformers等），并掌握TensorFlow或PyTorch等深度学习框架。能够独立完成项目实现，并具备较强的调试与解决问题的能力。

Job Responsibilities: Responsible for coding the project, including core tasks like data preprocessing, model building, and training.

Required Skills: Strong Python programming skills, familiarity with common NLP tools (like NLTK, spaCy, Transformers), and proficiency in deep learning frameworks such as TensorFlow or PyTorch. Must be able to independently implement projects and have strong debugging and problem-solving skills.

B成员：项目方向与理论支持

Project focus and theoretical foundation.

职责：负责收集和分析NLP的背景资料，梳理NLP发展历史，评估项目难度，明确关键任务并提供理论支撑，协助A成员优化技术方向。

所需能力：具备高效的文献调研和项目分析能力，能够提炼项目所需的NLP核心知识，并为项目的方向与模型选择提供策略性指导。需要良好的沟通表达能力，能够将分析结果清晰传达给团队。

Responsibilities include collecting and analyzing NLP background materials, reviewing the development history of NLP, evaluating project complexity, identifying key tasks, offering theoretical backing, and assisting team member A in refining the technical approach.

Required skills include efficient literature research and project analysis skills, the ability to distill essential NLP knowledge for the project, and offering strategic guidance on project direction and model selection. Strong communication skills are required, with the ability to clearly communicate analysis results to the team.

C成员：技术细节与工具研究

Research on technical details and tools.

职责：深入研读相关领域的技术论文，挖掘关键技术要点，并研究项目所需的工具库和模型实现方法，为A成员提供具体的技术支持。

所需能力：具备强大的文献阅读与信息提炼能力，熟练掌握NLP相关工具（如Scikit-learn、Transformers等）的具体用法，快速理解论文中的先进方法，并对项目的实现技术细节提供指导。

Responsibilities: Conduct in-depth research on technical papers in relevant fields, identify key technical points, and research the necessary tools and model implementation methods to provide specific technical support to team member A.

Required skills: Have strong skills in reading literature and extracting information, proficient in using NLP-related tools (such as Scikit-learn, Transformers, etc.), quickly grasp advanced methods presented in papers, and offer guidance on technical details for project implementation.

D成员：模型评估与优化建议

Model assessment and optimization recommendations.

职责：负责模型性能的分析与评估，搜集类似项目的优化案例，提出项目的改进方案，并为模型效果的持续优化提供建议。

所需能力：掌握模型评估的关键指标（如准确率、F1分数等）及分析方法，具备数据解读能力，能够识别模型的优势与不足。需具有较强的创新思维，能够结合案例提供适合本项目的优化策略。

Responsibilities: Analyze and evaluate model performance, gather optimization examples from similar projects, suggest project improvements, and provide ongoing optimization advice.

Required skills: Understand key model evaluation metrics (like accuracy, F1 score) and analysis methods, and have the ability to interpret data and identify model strengths and weaknesses. Must be highly innovative and able to develop optimization strategies tailored to this project based on case studies.

**3. 选择了这个选题之后，在做需求分析、后续收集数据时是否遇到过困难？是否修改过选题？**

首先，由于这个项目的起源较早，因此在进行NLP情感分析的优化时，我们面临着难以超越当前市场水平的挑战。近年来，市场上的工具和技术迅速发展，许多现有的情感分析模型和算法已经具备相当高的性能和准确性。这使得我们在追赶和提升模型效果方面感到力不从心，很难取得显著的技术突破。

First, since this project started quite early, we're finding it challenging to optimize NLP sentiment analysis to surpass current market standards. In recent years, market tools and technologies have evolved rapidly, with many existing sentiment analysis models and algorithms already performing at a high level. This makes it hard for us to catch up and significantly improve our model's effectiveness.

其次，作为一个底层技术项目，其后续应用的开发需要我们自行创建应用对接。这不仅增加了项目的复杂性，还使得整个实施过程变得更加繁琐。为了实现应用的顺利对接，我们必须投入额外的时间和精力来设计和开发适配方案，以确保最终系统的可用性和连贯性。这种额外的开发工作对团队提出了更高的要求，也增加了项目周期的不确定性。

Secondly, being a foundational technology project, we need to develop application interfaces ourselves for subsequent applications. This not only adds complexity to the project but also makes the implementation process more cumbersome. To ensure smooth integration, we have to put in extra time and effort to design and develop adaptation solutions, which increases the demands on our team and adds uncertainty to the project timeline.

此外，数据的收集和整理也成为一大难题。由于情感分析任务对数据的质量和多样性有着较高的要求，我们在数据源的选择和数据清洗过程中遇到了很多挑战。例如，部分数据集存在严重的不平衡问题，某些情感类别的数据量远远少于其他类别，导致训练模型时出现偏差。为了解决这些问题，我们不得不对数据进行采样、扩充和清洗，进一步增加了数据处理的工作量和复杂性。

Moreover, data collection and organization have become significant challenges. Sentiment analysis tasks require high-quality and diverse data, and we've faced numerous issues in selecting data sources and cleaning data. For instance, some datasets are highly imbalanced, with certain sentiment categories having far fewer samples than others, leading to bias during model training. To address these issues, we need to sample, augment, and clean the data, which further increases the workload and complexity of data processing.

我们曾考虑过更换选题，但由于已经投入了大量的前期准备工作，包括需求分析和数据收集，若此时更换选题将导致之前的努力白费，项目进度也会大幅度延误。因此，经过权衡，我们决定坚持原选题，并通过优化现有方案来克服遇到的困难。

We considered changing the project topic, but given the substantial preliminary work already done, including requirement analysis and data collection, switching topics now would render our previous efforts useless and significantly delay the project progress. Therefore, after weighing our options, we've decided to stick to the original topic and overcome the challenges by optimizing our current solution.

**4. 你们是如何搜寻数据集？如何整理数据集的？整理过程中遇到过什么困难？**

在搜寻数据集的过程中，我们最初考虑使用爬虫技术进行数据抓取，但由于这是我们首次尝试编写此类项目，抓取的数据在后续的数据处理、特征提取等方面可能会带来较大的复杂性。如果操作不当，不仅会影响模型训练的整体进度，甚至可能需要推翻重做。因此，我们决定放弃爬虫方案，转而使用现成的数据集。

While searching for datasets, we initially considered using web scraping to extract data. However, since this was our first time attempting such a project, the scraped data could introduce significant complexity in later stages of data processing and feature extraction. If not handled properly, it could affect the progress of model training and might even require us to start over. Therefore, we decided to abandon the web scraping plan and instead use existing datasets.To find suitable datasets, we searched multiple NLP data resource platforms, such as GitHub, Kaggle, Tianchi, Huggingface, CLUEDatasetSearch, and Datahub. After screening, we ultimately chose Weibo comments from the 2020 pandemic period for the first version of model training.

为了找到合适的数据集，我们在多个NLP数据资源平台上进行了搜索，如 GitHub、Kaggle、天池、Huggingface、CLUEDatasetSearch 和 Datahub 数据中心等。经过一番筛选，我们最终选择了2020年疫情期间的微博评论数据，用于进行第一版模型的训练。

在整理数据集的过程中，我们遇到了一些挑战。首先，数据集本身包含大量噪声，例如错别字、不完整的句子和重复内容，需要进行充分的数据清理。此外，数据标签的标准化也是一项重要的工作，确保情感分类的一致性和准确性。这些困难虽然增加了整理数据集的工作量，但通过逐步的清理和优化，我们得以获得更高质量的输入数据，以支持模型的训练。

While organizing the dataset, we encountered some challenges. First, the dataset had a lot of noise, such as typos, incomplete sentences, and duplicate content, which required thorough data cleaning. Additionally, standardizing the data labels was important to ensure the consistency and accuracy of sentiment classification. Although these difficulties increased the workload of organizing the dataset, through gradual cleaning and optimization efforts, we managed to obtain higher quality input data to support model training.

**5. 是否在整理的过程中更换过数据集？如何更换？**

我们一开始是打算模拟实战数据集，因此选择了一个包含十万条疫情期间微博评论的数据集。但是在后续处理中，我们发现由于数据集过大，训练速度非常缓慢，影响了项目进度。此外，由于数据集来自疫情期间，消极类别的数据远超其他两类，导致模型在后期对消极情感的准确率较高，这可能存在过拟合的倾向。而且由于数据集的时间较早，一些现代网络用语无法被正确分类。因此，我们决定更换数据集。

Initially, we intended to simulate real-world data conditions, so we chose a dataset of 100,000 Weibo comments from the pandemic period. However, we later found that the dataset's large size slowed down training, delaying the project. Additionally, the negative data was much more prevalent than the other categories, suggesting overfitting, and outdated internet terms weren't classified correctly. Therefore, we decided to switch datasets.

更换过程中，我们重新评估了数据需求，选择了一个较新且较为干净的酒店评论数据集。新的数据集规模适中，并且情感类别分布较为均衡，有助于减少过拟合的风险。我们对新数据集进行了相应的数据清理和标签标准化处理，以确保数据质量符合模型训练的要求。通过这一更换，我们得以改进模型的训练效果，提升情感分类的准确性。

We reassessed our data needs and selected a newer, cleaner hotel review dataset. This dataset is moderately sized with a balanced distribution of sentiment categories, reducing the risk of overfitting. We cleaned and standardized the data to ensure it met the training requirements. This switch improved our model's training effectiveness and sentiment classification accuracy.

6. 是否已经开始训练？训练效果展示。

我们已经开始了模型的训练并且已经取得了一定的进展，如图

已经训练了成品模型，正确率已经取得了可人的成绩，大概在90%左右

如图

结果显示**BERT-Chinese**模型在准确率、精确率、召回率和F1分数等各项指标上表现最佳，具有显著的优势；而**Word2Vec + SVM**表现最差，尤其在F1分数上较为逊色。整体而言，BERT模型在中文文本分类任务中具有较高的优越性。

We have begun training the model and have made some progress, as shown in the figure. We have trained the model to completion, achieving an accuracy of around 90%.

The results, as shown in the figure, indicate that the BERT-Chinese model outperforms others in terms of accuracy, precision, recall, and F1 score, demonstrating significant advantages. In contrast, the Word2Vec + SVM model shows the poorest performance, particularly in the F1 score. Overall, the BERT model proves to be highly superior for Chinese text classification tasks.

7. 接下来的具体时间安排计划

在接下来的时间里，我们打算继续优化我们的模型，使其获得更高的正确率，我们会更换现有代码中一些已经过时的实现方式，尝试采用更先进的方式实现他们，我们还考虑为这个模型提供多语言支持和emoji以及颜文字支持，使其更加的泛用。另外，介于这项技术能够实现很多广泛的应用，我们会尝试进行基于成品模型，进行应用的实现开发，对接现有成型体系，研究中间件，使我们的项目更加有用。

Moving forward, we aim to enhance our model's accuracy through continuous optimization. This includes updating outdated code with more advanced implementation methods. We're also planning to expand the model's capabilities by adding support for multiple languages, emojis, and kaomoji, making it more universally applicable. Furthermore, recognizing the broad potential applications of this technology, we intend to develop practical applications based on our refined model. This will involve integrating with existing systems, developing middleware solutions, and overall increasing the practical value of our project.

Our development roadmap includes:

WebUI开发

设计直观的用户界面，允许用户输入文本进行情感分析

实现结果可视化，如情感分布饼图或条形图

添加批量分析功能，支持上传CSV文件进行批量情感分析

实现用户反馈机制，收集错误分类的样本以进一步改进模型

WebUI Development:Creating an intuitive user interface for text-based sentiment analysisImplementing visual representations of results, such as sentiment distribution chartsAdding functionality for batch analysis, including CSV file upload supportDeveloping a user feedback system to collect misclassified samples for model improvement

LLM集成

研究如何将我们的情感分析模型与大型语言模型(如GPT)结合

探索利用LLM进行零样本或少样本学习，以提高模型在新领域的泛化能力

Large Language Model (LLM) Integration:Exploring ways to combine our sentiment analysis model with LLMs like GPTInvestigating zero-shot or few-shot learning techniques to improve the model's adaptability to new domains

Voice功能开发

集成语音识别技术，允许用户通过语音输入进行情感分析

开发语音合成功能，根据文本的情感分析结果生成相应情感的语音

Voice Feature Development:Incorporating speech recognition for voice-based sentiment analysis inputDeveloping speech synthesis capabilities to generate emotionally appropriate voice outputs based on sentiment analysis resultsThese enhancements will significantly broaden the application scope of our sentiment analysis tool, making it more robust, user-friendly, and adaptable to various use cases.