**Background:** Music, as an ancient art form profoundly influencing human emotions, plays a central role in various cultures and societies and is vital in our emotional expression and experience. Whether in moments of celebration or sorrow, music touches our hearts in unique ways.

**Significance:** Considering music's significant role in influencing individual emotions and mental states, my project aims to deeply understand the emotions conveyed through music by its lyrics content and audio features. By proving the effectiveness of a comprehensive model (combining lyrics and audio features), I hope to help people better understand the emotions experienced through music, thereby improving emotional health and potentially alleviating mental health issues like anxiety and depression.

**Related Solution (Research):** Music Emotion Recognition (MER) is a key research direction in the field of Music Information Retrieval, utilizing lyrics and audio features to predict the emotional attributes of music. Inspired by the Russell emotion model, the MER field has already made more detailed emotion classifications. Building on this, I plan to develop a more comprehensive and accurate music emotion recognition composite model, focusing on both lyrics and audio aspects. In my literature review, I have selected this paper as a baseline, which holds significant standing in the MER field for employing Bi-directional Long Short-Term Memory (Bi-LSTM) deep learning methods combined with GloVe word representation weights, achieving an accuracy of 91.08% in emotion classification using song lyrics. My goal is to surpass this benchmark, achieving progress in the refinement, accuracy, and general applicability of emotion classification through a composite model.

**Methodology:**

**Dataset:** In my research project, I used two datasets related to music emotion, both provided by the same author. The first dataset, "MoodyLyrics," consists of songs annotated according to the four quadrants of the Russell emotion model based on their lyric text (labels **only from lyrics)**. The second dataset, "MoodyLyrics4Q," includes songs tagged as one of the four categories of the Russell model based on tags from Last.fm(labels from **overall music tags**).

Due to copyright issues, the original datasets did not provide lyrics, so I had to obtain them myself. Initially, I attempted to match lyrics using the Genius API (lyricsgenius) based on song names and artist names. However, I quickly realized this method could lead to incorrect matches when song titles or artist names had multiple spellings. To improve accuracy, I switched to a custom web scraping approach. By searching Google and parsing HTML files from the Genius website, I was able to more precisely locate and verify song and artist information, effectively reducing the risk of retrieving incorrect lyrics.

Additionally, I used the official Spotify API to extract audio features of the songs, also based on song names and artist information. After completing these steps, I thoroughly cleaned and standardized the lyrics data scraped from the Genius website. I used custom regular expressions to remove non-essential information (like “[Verse]” tags) and filtered out non-English lyrics, ensuring the dataset's quality and relevance.

Finally, to balance Dataset 1, I applied downsampling techniques, randomly removing 90 songs labeled as "happy," using a specific random state to ensure reproducibility. Additionally, before initiating the training process, the entire dataset was randomly shuffled. This step is crucial to prevent the model from learning any potential order present in the data, ensuring it learns based on the actual features of the songs. PPT show结构

**Reproducing the paper**

After preparing the data, I began the process of replicating a key paper, an extremely important step in scientific research. The main purpose of replicating the paper was to verify the reliability and effectiveness of the original research results. This process not only helped me confirm the reproducibility of the research findings but also deepened my understanding of the original study's methods and logic. It also provided me with opportunities to identify and improve potential issues.

During the replication process, I strictly followed the hyperparameters and structure mentioned in the paper. Specifically, the paper used pretrained GloVe 100-dimensional vectors for word embedding. Since the Naive Bayes (NB) model does not accept negative values, I chose the TF-IDF method for word embedding testing in the NB model, while other models continued to use GloVe. Following the parameters set in the paper, I successfully replicated several models, including Naive Bayes, K-Nearest Neighbors, Support Vector Machine, Convolutional Neural Network, Long Short-Term Memory Network, and Bidirectional Long Short-Term Memory Network. Each model achieved an accuracy similar to that in the original paper.

**Experiment Design:**

As I delve deeper into the analysis of the original paper, I am now preparing to further expand my research. My goal is to enhance the performance and accuracy of our model by exploring different experimental methods. For this, I plan to conduct experiments from three main aspects: embedding technologies, preprocessing techniques, and the application of audio features.

**Embedding:**

In this process, I have identified and implemented several key improvements. Specifically, in dealing with the maximum sequence length (max\_len), I decided to reduce it from 1000, as used in the paper, to 250. This change is based on the simple fact that the actual length of all text sequences in the dataset is far less than 1000, which can reduce the impact of 0 padding (noise) on model performance.

After optimizing the maximum length of text sequences, I explored different word embedding technologies. I selected Bag of Words (BoW), Term Frequency-Inverse Document Frequency (TFIDF), and Word2Vec300d, and applied a uniform preprocessing strategy to each model, including lemmatization, lowercasing, noise removal, and stopword removal. These steps ensured the cleanliness and consistency of the data, laying a solid foundation for subsequent analysis.

In terms of model tuning, I finely adjusted key parameters of traditional machine learning models like Naive Bayes, Support Vector Machine, and K-Nearest Neighbors using a grid search strategy, such as n-gram range and document frequency, thus enhancing the model's cross-validation performance. For deep learning models like Text-CNN and BiLSTM, I optimized the model's generalization ability by analyzing learning and loss curves, adjusting network structure, neuron numbers, dropout rates, and learning rates.

My experimental results indicate that BoW, TFIDF, and Word2Vec embedding technologies demonstrate superior performance in processing lyrics data compared to the GloVe used in the original paper, surpassing the paper's baseline accuracy of 91. BoW and TFIDF, which are based on word frequency and bag of words structures, effectively highlighted key words and phrases, while Word2Vec's CBOW and Skip-gram models focused on capturing the relationship between vocabulary and its context, offering a more nuanced perspective for understanding emotions in lyrics.

After further exploring four core text classification models and their corresponding word embedding technologies, I found that each had unique advantages. For example, Naive Bayes combined with the bag of words model excels in handling high-dimensional data, particularly suitable for scenarios where keyword frequency is higher than context. Support Vector Machine combined with TFIDF effectively combines SVM's ability to optimize classification boundaries in high-dimensional spaces with TFIDF's emphasis on the uniqueness of important words, making it an ideal choice for processing multi-topic texts. Text-CNN combined with Word2Vec uses convolutional layers to capture local features, and the deep word embeddings provided by Word2Vec help process semantically rich texts. Finally, the combination of Bi-directional Long Short-Term Memory Network and Word2Vec excels in understanding long-distance dependencies and complex semantics, offering a comprehensive text understanding by capturing forward and backward dependencies in the text.

**Preprocessing:**

To understand the impact of different preprocessing strategies on model performance, I conducted a series of experiments on the four best-performing models from the embedding stage (including Naive Bayes combined with the bag of words model, Support Vector Machine with TFIDF, Text-CNN combined with Word2Vec, and Bi-directional Long Short-Term Memory Network with Word2Vec). These experiments aimed to explore the specific impact of different preprocessing methods like Stemming, Lemmatization, Noise Removal (NR), and Stopword Removal (SR) on model performance.

To ensure the stability and reliability of experimental results, especially for deep learning models like Text-CNN and BiLSTM, I designed a loop testing method. Considering that these models start with random weight initialization during training, I conducted at least three tests on them to ensure the obtained results were stable and credible.

After finalizing the embedding models, tuning, and preprocessing methods, I observed a significant milestone: On individual lyric analysis, models like SVM, Naive Bayes, and Text-CNN have already surpassed the baseline accuracy established in the original paper. Notably, the SVM model achieved an impressive accuracy and F1 score of 94%

**Audio Features:**

In the third phase of my research, I focused on exploring the impact of audio feature fusion on model performance. Before starting, I first normalized the audio features to avoid changes in the original units and distribution of the data that might be caused by MinMax normalization. This step provided a consistent and reliable data foundation for subsequent data analysis and model training.

Then, using visualization tools such as heat maps, PCA, and t-SNE, as well as the Hopkins statistic, I delved into understanding the distribution and clustering potential of audio features in the data space. The Hopkins test results showed some degree of clustering tendency. However, I noticed considerable randomness in the data points during PCA and t-SNE dimensionality reduction, which might suggest a loss of correlation in high-dimensional data during reduction. Particularly in the analysis of heat maps, I found that audio features did not show strong positive or negative correlations with emotional categories. For example, although there was a certain positive correlation between energy and loudness in the Angry category, other audio features did not show significant correlations.

While exploring the impact of audio features like key, mode, and time signature, I found through experiments with heat maps and random forest models that these features contributed little to emotion classification. Due to potential issues of the curse of dimensionality, I decided to exclude these features from subsequent model training. Additionally, I conducted separate training and prediction with data containing only audio features, but the results were not ideal.

Next, I integrated audio features into previously well-performing models for in-depth experiments. I tried a feature pre-fusion method with SVM, combining text and audio features into one vector; in the Text-CNN model, I added an audio feature input layer and combined text and audio features through a merging layer; a similar method was used in the BiLSTM model. I also explored a stacking-based ensemble learning approach, where I chose the random forest, which performed best in the audio feature experiments, to handle audio features, TFIDF and SVM for text features, and XGBoost as the meta-classifier.

Although audio features did not significantly improve model performance in preliminary experiments, I speculated that this might be because the dataset used primarily focused on the emotional dimension of lyrics, with limited influence of audio features. Therefore, I decided to conduct performance tests using the second dataset to compare the performance of single models and composite models. The test results showed that although the overall accuracy was low, possibly due to the presence of many unseen tokens in the test set, the composite models outperformed the single models in terms of F1 scores, especially in the CNN model, where the composite model's F1 score reached 38%, compared to a maximum of 36% for single models, with the Angry category improving by 2-5% in F1 scores.

To further validate the effectiveness of audio features, I trained and tested on the second dataset, using a paper employing XL-NET and Lemmatization for emotion classification as a benchmark. This study achieved an F1 score of 59% on this dataset. In comparison, my composite model far exceeded single models in F1 scores, with the CNN model reaching up to 67%, and all composite models' F1 scores surpassing single models and exceeding the paper's benchmark.

The visualization analysis on Dataset 2 revealed strong positive and negative correlations between audio features and emotional categories, and PCA analysis also showed certain clustering patterns. These results proved that considering both lyrics and audio features could achieve more accurate predictions in emotion classification tasks, enhancing the model's generalization ability.

**Use Case：**

After completing an in-depth analysis of audio and text features, I decided to apply the model I developed to a practically challenging scenario – emotion analysis of Spotify's Top 100 songs over the past decade (2013 to 2023). The aim was to verify the potential and accuracy of my model in handling real-world data.

In choosing the model, I based my decision on cross-validation results and test set performance on the two datasets. After careful comparison and analysis, I selected the CNN model trained on Dataset 2, as it not only demonstrated the best performance but also had excellent generalization capabilities.

My prediction results revealed some notable trends. The number of songs marked as "Sad" significantly increased in both 2020 and 2022, likely related to the global COVID-19 pandemic in those years. Moreover, from 2020 to 2022, the number of songs classified as "Happy" continually decreased, reaching its lowest point in the last five years by 2022. This downward trend might reflect the impact of global events, such as the Russia-Ukraine conflict in 2022, which could have triggered widespread unrest and negative emotions worldwide.

**Evaluation and Discussion:**

In my journey of exploring Music Emotion Recognition (MER), I first ensured the quality of the datasets. I used two datasets provided by the same author, "MoodyLyrics" and "MoodyLyrics4Q," which contain songs annotated according to the Russell emotion model. Since the original datasets did not provide lyrics, I adopted a two-stage method to obtain them. Initially, I tried using the Genius API for matching, but due to the challenge of spelling diversity, I switched to a custom web scraping method to improve match accuracy. I ensured the cleanliness and consistency of the dataset through normalization and cleaning processes, and ultimately achieved data balance using downsampling techniques. Additionally, to further enhance the dataset's robustness for training, I implemented random shuffling of the entire dataset. These meticulous preparations laid a solid foundation for subsequent experiments.

Replicating the paper was a process that verified the reliability of the original study and deepened my understanding of research methods and model functions. Strictly following the parameters of the original paper, I not only replicated multiple models but also gained insights into the effectiveness of the combination of Bi-directional Long Short-Term Memory networks (BiLSTM) with GloVe. I also identified potential issues in the paper and proposed experiments to address them.

In the experiment design, I focused on word embedding techniques, preprocessing techniques, and the application of audio features. My experiments showed that shortening the maximum sequence length significantly improved model performance while reducing computational costs. In terms of word embedding, I found BoW, TFIDF and Word2Vec to perform better than GloVe when dealing with repetitive and rhythmic text such as lyrics, which highlights the importance of choosing the right word embedding technique based on the data and task characteristics.

In model tuning, I gained a deep understanding of the strengths and functions of each model. I chose the appropriate lyric model and used global searches, learning validation loss curves, and confusion matrices for model evaluation and tuning. In evaluating preprocessing techniques, I found that appropriate preprocessing strategies significantly improved the accuracy and efficiency of the model, emphasizing the necessity of fine preprocessing.

The integration assessment of audio features initially showed that while they did not significantly enhance model performance in Dataset 1 (MoodyLyrics), which is solely based on lyrics, they contributed notably to model accuracy in more extensive tests. This became particularly evident when Dataset 2 (MoodyLyrics4Q) was used as a test set to evaluate the models trained on Dataset 1. The models that combined both lyrics and audio features (multimodal models) outperformed those relying on a single modality, demonstrating the value of a comprehensive analysis approach in enhancing accuracy and generalizability. Further, when Dataset 2 was used for both training and testing, the multimodal models not only significantly surpassed the performance of the unimodal models but also exceeded the baseline accuracy of Dataset 2. This strongly supports the effectiveness of incorporating both audio and lyrical features in music analysis, underscoring the comprehensive understanding they provide compared to a single modality approach.

In practical application, I applied the model to the emotion analysis of Spotify's Top 100 songs over the past ten years, choosing the best-performing CNN model based on cross-validation and tests on two datasets. Using significant world events as a validation point, this not only verified the model's application potential but also demonstrated its accuracy and generalizability in real-world data. The dataset can be provided for further research by other studies.

In conclusion, my project was meticulously designed and implemented at every stage. Despite challenges, each phase provided important insights and helped me continuously optimize the model. Particularly in the areas of word embedding techniques and audio features, my experiments revealed the potential to enhance the performance of music emotion analysis models.

Next step

Conduct a detailed emotional analysis of Spotify's Top 100 songs annually.

Develop and finalize a comprehensive report on the findings.

背景：音乐，作为一种古老的艺术形式，深刻地影响着人类的情感，在不同文化和社会中扮演着核心角色，对我们的情感表达和体验至关重要。无论是在庆祝的时刻还是悲伤的时候，音乐都以独特的方式触动我们的心灵。

重要性：考虑到音乐在影响个体情绪和心理状态方面的重要作用，我的项目旨在深入理解通过音乐的歌词内容和音频特征传达的情感。通过验证一个综合模型（结合歌词和音频特征）的有效性，我希望帮助人们更好地理解通过音乐体验的情感，从而改善情感健康，并有可能缓解像焦虑和抑郁这样的心理健康问题。

相关解决方案（研究）：音乐情感识别（MER）是音乐信息检索领域的一个关键研究方向，它利用歌词和音频特征来预测音乐的情感属性。受到罗素情感模型的启发，MER领域已经做出了更详细的情感分类。在此基础上，我计划开发一个更全面、更准确的音乐情感识别综合模型，关注歌词和音频方面。在我的文献回顾中，我选择了这篇论文作为基线，该论文在MER领域具有重要地位，它采用双向长短期记忆（Bi-LSTM）深度学习方法结合GloVe词表示权重，在使用歌词进行情感分类方面达到了91.08%的准确率。我的目标是超越这个基准，实现情感分类的精细化、准确性和普遍适用性方面的进步。

方法论： 数据集：在我的研究项目中，我使用了两个与音乐情感相关的数据集，均由同一作者提供。第一个数据集“MoodyLyrics”包含根据歌词文本（仅基于歌词的标签）按照罗素情感模型的四个象限标注的歌曲。第二个数据集“MoodyLyrics4Q”包括根据Last.fm的标签（基于整体音乐标签）划分为罗素模型四个类别之一的歌曲。由于版权问题，原始数据集没有提供歌词，所以我不得不自己获取。最初，我尝试使用Genius API（lyricsgenius）根据歌名和艺术家名匹配歌词。然而，我很快意识到，当歌曲标题或艺术家名称有多种拼写时，这种方法可能导致错误匹配。为了提高准确性，我转向了自定义的网络爬虫方法。通过搜索Google并解析来自Genius网站的HTML文件，我能够更准确地定位并验证歌曲和艺术家信息，有效地减少了检索错误歌词的风险。 此外，我还使用了官方Spotify API提取歌曲的音频特征，也是基于歌曲名称和艺术家信息。完成这些步骤后，我彻底清洗和标准化了从Genius网站爬取的歌词数据。我使用自定义的正则表达式来移除非必要信息（如“[Verse]”标签）并过滤掉非英文歌词，确保数据集的质量和相关性。 最后，为了平衡数据集1，我应用了降采样技术，随机移除了90首标记为“快乐”的歌曲，使用特定的随机状态以确保可重复性。此外，在开始训练过程之前，整个数据集被随机洗牌。这一步骤对于防止模型学习数据中可能存在的任何顺序至关重要，确保它基于歌曲的实际特征进行学习。PPT展示结构

**论文复现：** 在数据准备完毕后，我开始了复现关键论文的过程，这是科学研究中极为重要的一步。复现论文的主要目的是验证原始研究结果的可靠性和有效性。这一过程不仅帮助我确认了研究发现的可复现性，而且加深了我对原始研究方法和逻辑的理解，并为我识别和改进潜在问题提供了机会。

在复现过程中，我严格遵循论文中提到的超参数和结构。具体来说，该论文使用预训练的GloVe 100维向量进行词嵌入。由于朴素贝叶斯（NB）模型不接受负值，我选择了TF-IDF方法用于NB模型的词嵌入测试，而其他模型继续使用GloVe。遵循论文设定的参数，我成功复现了多个模型，包括朴素贝叶斯、K最近邻、支持向量机、卷积神经网络、长短期记忆网络和双向长短期记忆网络。每个模型都达到了与原始论文相似的准确率。

**实验设计：** 随着我对原始论文的分析越来越深入，我现在准备进一步扩展我的研究。我的目标是通过探索不同的实验方法，提高模型的性能和准确率。为此，我计划从三个主要方面进行实验：嵌入技术、预处理技术和音频特征的应用。

**嵌入：** 在这个过程中，我确定并实施了几项关键改进。具体来说，在处理最大序列长度（max\_len）时，我决定将其从论文中使用的1000减少到250。这一改变基于一个简单的事实，即数据集中所有文本序列的实际长度远小于1000，这可以减少0填充（噪声）对模型性能的影响。

在优化文本序列的最大长度后，我探索了不同的词嵌入技术。我选择了词袋（BoW）、词频-逆文档频率（TFIDF）和Word2Vec300d，并对每个模型应用了统一的预处理策略，包括词形还原、小写化、噪声移除和停用词移除。这些步骤确保了数据的清洁和一致性，为后续分析奠定了坚实的基础。

在模型调优方面，我使用网格搜索策略对传统机器学习模型（如朴素贝叶斯、支持向量机和K最近邻）的关键参数进行了精细调整，例如n-gram范围和文档频率，从而提高了模型的交叉验证性能。对于像Text-CNN和BiLSTM这样的深度学习模型，我通过分析学习和损失曲线，调整网络结构、神经元数量、dropout率和学习率，优化了模型的泛化能力。

我的实验结果表明，BoW、TFIDF和Word2Vec嵌入技术在处理歌词数据方面的性能优于原始论文中使用的GloVe，超过了论文的基线准确率91。BoW和TFIDF基于词频和词袋结构，有效突出了关键词和短语，而Word2Vec的CBOW和Skip-gram模型专注于捕捉词汇及其上下文之间的关系，为理解歌词中的情感提供了更细腻的视角。

在进一步探索四种核心文本分类模型及其相应的词嵌入技术后，我发现每种模型都有其独特优势。例如，朴素贝叶斯结合词袋模型擅长处理高维数据，特别适合于关键词频率高于上下文的情景。支持向量机结合TFIDF有效地结合了SVM在高维空间优化分类边界的能力与TFIDF强调重要词的独特性，使其成为处理多主题文本的理想选择。Text-CNN结合Word2Vec使用卷积层捕捉局部特征，而Word2Vec提供的深度词嵌入有助于处理语义丰富的文本。最后，双向长短期记忆网络和Word2Vec的结合擅长理解长距离依赖和复杂语义，通过捕捉文本中的前向和后向依赖，提供全面的文本理解。

**预处理：** 为了了解不同预处理策略对模型性能的影响，我对嵌入阶段表现最佳的四种模型（包括与词袋模型结合的朴素贝叶斯、TFIDF支持向量机、与Word2Vec结合的Text-CNN和双向长短期记忆网络）进行了一系列实验。这些实验旨在探索不同预处理方法，如词干提取、词形还原、噪声移除（NR）和停用词移除（SR）对模型性能的具体影响。

为了确保实验结果的稳定性和可靠性，特别是对于像Text-CNN和BiLSTM这样的深度学习模型，我设计了一种循环测试方法。考虑到这些模型在训练过程中从随机权重初始化开始，我对它们进行了至少三次测试，以确保获得的结果稳定且可信。

在确定了嵌入模型、调优和预处理方法后，我观察到一个重要的里程碑：在单独的歌词分析上，像SVM、朴素贝叶斯和Text-CNN这样的模型已经超越了原始论文中建立的基线准确率。值得注意的是，SVM模型达到了令人印象深刻的准确率和F1分数94%。

**音频特征：** 在我的研究的第三阶段，我专注于探索音频特征融合对模型性能的影响。在开始之前，我首先对音频特征进行了标准化，以避免因MinMax标准化而对原始单位和数据分布的改变。这一步为后续数据分析和模型训练提供了一致且可靠的数据基础。

然后，我使用热图、PCA和t-SNE等可视化工具，以及霍普金斯统计量，深入理解音频特征在数据空间中的分布和聚类潜力。霍普金斯测试结果显示了一定程度的聚类倾向。然而，我在PCA和t-SNE降维过程中注意到数据点存在相当大的随机性，这可能表明在降维过程中高维数据丢失了相关性。特别是在分析热图时，我发现音频特征与情感类别之间没有表现出强烈的正相关或负相关。例如，虽然在愤怒类别中能量和响度之间存在一定的正相关性，但其他音频特征并没有显示出显著的相关性。

在探索音频特征如调性、模式和节拍签名的影响时，我通过热图和随机森林模型的实验发现，这些特征对情感分类的贡献很小。由于维数灾难的潜在问题，我决定在后续模型训练中排除这些特征。此外，我还单独对只包含音频特征的数据进行了训练和预测，但结果并不理想。

接下来，我将音频特征整合到之前表现良好的模型中进行深入实验。我尝试了与SVM的特征预融合方法，将文本和音频特征合并成一个向量；在Text-CNN模型中，我添加了一个音频特征输入层，并通过合并层将文本和音频特征结合起来；BiLSTM模型也采用了类似的方法。我还探索了基于堆栈的集成学习方法，选择在音频特征实验中表现最佳的随机森林处理音频特征，TFIDF和SVM处理文本特征，以及XGBoost作为元分类器。

尽管在初步实验中音频特征并没有显著提高模型性能，但我推测这可能是因为所使用的数据集主要关注歌词的情感维度，音频特征的影响有限。因此，我决定使用第二个数据集进行性能测试，比较单一模型和复合模型的性能。测试结果显示，尽管总体准确率较低，可能是因为测试集中出现了许多未见过的标记，但在F1分数方面，复合模型优于单一模型，特别是在CNN模型中，复合模型的F1分数达到了38%，而单一模型的最高分数为36%，愤怒类别的F1分数提高了2-5%。

为了进一步验证音频特征的有效性，我在第二个数据集上进行了训练和测试，以一篇使用XL-NET和词形还原进行情感分类的论文作为基准。该研究在这个数据集上实现了59%的F1分数。相比之下，我的复合模型在F1分数上远超单一模型，CNN模型达到了67%，所有复合模型的F1分数都超过了单一模型，并超过了论文的基准。

第二个数据集的可视化分析显示，音频特征与情感类别之间存在强烈的正负相关性，PCA分析也显示了一定的聚类模式。这些结果证明，考虑歌词和音频特征可以在情感分类任务中实现更准确的预测，提高模型的泛化能力。

**实际用例：** 在深入分析音频和文本特征之后，我决定将我开发的模型应用于一个具有实际挑战性的场景——对过去十年（2013至2023年）Spotify排行榜前100首歌曲的情感分析。目的是验证我的模型在处理现实世界数据方面的潜力和准确性。

在选择模型时，我基于对两个数据集的交叉验证结果和测试集表现进行了决策。经过仔细比较和分析，我选择了在数据集2上训练的CNN模型，因为它不仅展现了最佳性能，而且具有出色的泛化能力。

我的预测结果揭示了一些显著趋势。在2020年和2022年，被标记为“Sad”（悲伤）的歌曲数量显著增加，这可能与那些年全球COVID-19大流行有关。此外，从2020年到2022年，被分类为“Happy”（快乐）的歌曲数量持续减少，到2022年达到过去五年的最低点。这一下降趋势可能反映了全球事件（如2022年的俄罗斯-乌克兰冲突）对全球范围内普遍不安和负面情绪的影响。

**评估和讨论：** 在探索音乐情感识别（MER）的过程中，我首先确保了数据集的质量。我使用了两个由同一作者提供的数据集，“MoodyLyrics”和“MoodyLyrics4Q”，这些数据集包含根据罗素情感模型注释的歌曲。由于原始数据集没有提供歌词，我采用了两阶段方法来获取它们。最初，我尝试使用Genius API进行匹配，但由于拼写多样性的挑战，我转向了自定义的网络爬虫方法以提高匹配准确性。我通过标准化和清洗过程确保了数据集的清洁和一致性，并最终通过降采样技术实现了数据平衡。此外，为了进一步增强数据集的健壮性，我实施了对整个数据集的随机洗牌。这些细致的准备为后续实验奠定了坚实的基础。

复现论文的过程验证了原始研究的可靠性，并加深了我对研究方法和模型功能的理解。我严格遵循原始论文的参数，不仅复现了多个模型，还深入了解了双向长短期记忆网络（BiLSTM）与GloVe结合的有效性。我还识别了论文中的潜在问题，并提出了解决这些问题的实验。

在实验设计中，我专注于词嵌入技术、预处理技术和音频特征的应用。我的实验表明，缩短最大序列长度显著提高了模型性能，同时减少了计算成本。在词嵌入方面，我发现BoW、TFIDF和Word2Vec在处理重复和节奏性文本（如歌词）时的表现优于GloVe，这突显了基于数据和任务特性选择正确的词嵌入技术的重要性。

在模型调优方面，我深入理解了每个模型的优势和功能。我选择了合适的歌词模型，并使用全局搜索、学习验证损失曲线和混淆矩阵进行模型评估和调优。在评估预处理技术时，我发现适当的预处理策略显著提高了模型的准确性和效率，强调了精细预处理的必要性。

音频特征的综合评估最初显示，尽管它们在仅基于歌词的数据集1（MoodyLyrics）中并没有显著提高模型性能，但在更广泛的测试中，它们对模型准确性的贡献值得注意。这在将数据集2（MoodyLyrics4Q）用作测试集以评估在数据集1上训练的模型时变得尤为明显。结合了歌词和音频特征的多模态模型表现优于仅依赖单一模态的模型，证明了综合分析方法在提高准确性和泛化能力方面的价值。进一步地，当使用数据集2进行训练和测试时，多模态模型不仅显著超越了单模态模型，而且超过了数据集2的基线准确率。这强烈支持了在音乐分析中同时纳入音频和歌词特征的有效性，强调了与单一模态方法相比它们提供的全面理解。

在实际应用中，我将该模型应用于过去十年Spotify排行榜前100首歌曲的情感分析，选择了基于两个数据集的交叉验证和测试表现最佳的CNN模型。通过使用重大世界事件作为验证点，这不仅验证了模型的应用潜力，还展示了其在现实世界数据中的准确性和泛化能力。该数据集可以提供给其他研究进行进一步研究。

**总结：** 我的项目在每个阶段都经过了精心设计和实施。尽管面临挑战，但每个阶段都提供了重要的见解，帮助我不断优化模型。特别是在词嵌入技术和音频特征方面，我的实验揭示了提高音乐情感分析模型性能的潜力。

**下一步计划：**

1. 对Spotify排行榜前100首歌曲进行年度详细的情感分析。
2. 开发并完成对发现结果的综合报告。