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.