Background: (PPT 3)

As an ancient art form that profoundly influences human emotions, music plays a central role in various cultures and societies and is essential in our emotional expression and experience.

Significance:(PPT 4)

My project aims to disentangle the emotions conveyed by the lyrics and audio features of music. By developing a model that captures these elements, I hope to deepen our understanding of the emotional impact of music. This can enhance emotional well-being and may help address mental health issues such as anxiety and depression.

Research: (PPT 5)

My research focuses on music emotion recognition (MER), a field in music information retrieval. MER takes inspiration from the Russell emotion model and provides a more nuanced classification of emotions. My goal is to create a more advanced MER model that considers both lyrics and audio. I build on an important paper in the field that uses a Bi-directional Long Short-Term Memory (Bi-LSTM) model with GloVe word representations and achieves 91.08% accuracy in classifying sentiment from lyrics. My goal is to go beyond this benchmark and improve the accuracy and wider applicability of the model.

Method:

Datasets: (PPT 6-7)

My research uses two music sentiment datasets, "MoodyLyrics" and "MoodyLyrics4Q", both of which are divided into 4 categories based on the Russell sentiment model. The former uses only the emotional dimension of the lyric text as labels, while the latter uses the overall music labels of Last.fm. Due to copyright restrictions, I collected the lyrics myself, initially using the Genius API, but later designed a custom web crawl in order to be more accurate. This method ensures the correct match of song and artist and reduces the risk of wrong lyrics.

(PPT 7)

I also used the Spotify API to extract audio features from the songs. After collecting the data, I cleaned and standardized the lyrics, using custom regular expressions to eliminate unnecessary elements and filter out non-English lyrics.

(PPT 8-9)

To balance the first dataset, I randomly removed 90 songs from the "happy" category, keeping a specific random state to improve repeatability. Before training the model, I shuffled the entire dataset to avoid any bias in the original order of the data, ensuring that the model learned based on the actual characteristics of the song.

Reproducing the paper(PPT 10)

1.The next key stage of my project is to reproduce the research papers of Bi-LSTM and GloVe from the literature review, which is a fundamental aspect of scientific validation. This step is able to verify the reliability and validity of the original research results. It provides deeper insights into the methodology and logic of the original study, as well as identifying potential problems and directions for improvement.

2.In the reproduction phase, I strictly followed the hyperparameters and structure detailed in the paper. The study utilizes pre-trained GloVe 100-dimensional vectors for word embeddings.

3.However, for the Naive Bayes (NB) model, which cannot handle negative values, I chose the TF-IDF method for word embeddings, while the other models continued to use GloVe.
4.Following the guidelines of the paper closely, I was able to reproduce several models including Naive Bayes, K-nearest Neighbors, Support Vector machines, Convolutional neural networks, Long Short-Term Memory networks, and bidirectional long short-term memory networks, achieving similar accuracy as reported in the original paper.

Experimental design: (PPT 11)

Based on my analysis of the paper, I set out to improve the performance and accuracy of the model through different experimental methods. This involves three key areas: embedding techniques, preprocessing techniques, and incorporation of audio features.

Word embedding: (PPT 12)

1.I have made some improvements in this area. For example, I reduced the maximum sequence length (max_len) from 1000 to 250 based on the actual text sequence length in the dataset to minimize the effect of zero-padding. Then I realized that for repetitive and rhythmic texts such as lyrics, it may be effective to experiment with words and smaller contexts.
2.So next I tried various word embedding techniques, including Bag of Words (BoW), Term Frequency-Inverse Document Frequency (TFIDF), and Word2Vec300d, setting up a unified preprocessing step, lemma plus lowercase plus noise removal, and stop word removal. This ensures that the data is clean and consistent.
3.In addition, the parameters of the machine learning model and the deep learning model are fine-tuned to optimize their performance.
4.Experimental results show that BoW, TFIDF, and Word2Vec outperform GloVe, exceeding the baseline accuracy of 91% of the original paper.

Preprocessing: (PPT 13)

1.I experimented with the four top performing models using different preprocessing strategies from the embedding stage.
2.This includes exploring the impact of techniques such as stemming, lemmatization, noise cancellation, and stop word elimination.
3.For deep learning models such as Text-CNN and BiLSTM, I adopt a cyclic testing approach to account for random weight initialization during training to ensure stable and credible results.
4.With these approaches, I have achieved an important milestone: models like SVM, Naive Bayes, and Text-CNN have surpassed the baseline accuracy of the original paper in single lyrics analysis. **(PPT 14)** Notably, the SVM model achieved an impressive accuracy and F1 score of 94%. This phase is crucial and demonstrates the potential of improved embedding and preprocessing techniques to improve the accuracy of lyric emotion recognition.

Audio feature exploration: (PPT 15)

1.After studying word embeddings, tuning, and preprocessing, I focused on studying the impact of audio features on sentiment. Initially, I was prepared to understand the data first. Due to the particularities of the data, I standardized the audio features to maintain the consistency of the data, which is crucial for model training and analysis.

2.My analysis started by visualizing the distribution and clustering of audio features, using tools such as heatmaps, PCA(principal component Analysis), and t-SNE (t-distributed Random Neighborhood embedding), to assess the clustering trends by the Hopkins statistic. Although the Hopkins test implies some clustering, dimensionality reduction techniques such as PCA and t-SNE show considerable randomness in the data points. **3**.Heatmap analysis showed limited correlation between audio features and emotion categories, but among the audio features, energy and loundess had the strongest positive correlation with the angry class but the other features did not have strong positive or negative correlation, indicating the complexity of directly associating these features with specific emotions.

4.Subsequent investigations using heatmaps and random forest models on specific audio features such as keys, patterns, and temporal signatures showed minimal contribution to sentiment classification. **5.**Given these findings, I excluded these features from future model training. Models trained only with audio features did not produce satisfactory results.
(PPT 16) DATASET1 数据可视化展示

(PPT 17)
6.Integrating audio features: I then incorporated audio features into the previously successful model. This includes pre-fusing features for the SVM model and adding an audio feature input layer to the Text-CNN and BiLSTM models. I also experimented with stack-based ensemble learning methods, using random forest for audio features, TFIDF and SVM for text, and XGBoost as a meta-classifier.

7.Subsequent experiments show that audio features do not significantly improve the performance of the model. It is possible that this is because Dataset 1 is labeled according to the lyric sentiment dimension and the audio features do not have a very strong influence.

8.A subsequent performance test using the second dataset confirmed this, as the integrated model generally outperformed the unimodal model, especially in the CNN model, where the integrated model achieved an F1 score of 38%. The f1 improvement of angry class is obvious. This may be due to the positive correlation between energy and loundess of dataset 1 and angry class, which is learned by the model. This proves that considering both audio and lyrics can improve the generalization of the model.

9.To further validate my findings, I trained the composite model on dataset 2 and compared it to a benchmark study using XL-NET and lemma. My model outperformed the F1 of benchmark 59, achieving an F1 score of 67%. We also demonstrate in one step the value of fusing lyrics and audio features for sentiment classification.

10. I found a strong correlation between audio features and emotion categories in dataset 2; The PCA and heatmap results show the clustering and correlation between audio features and sentiment, which verifies the effectiveness of the fusion of lyrics and audio features for sentiment classification

(PPT 18) DATASET2 数据可视化展示
(PPT 19)综合 CNN Architecture 展示

Use case: (PPT20)

1.I then used my model to analyze sentiment in the top 100 Spotify songs from the last decade. This real-world application aims to validate the model's potential and generalizability.

2.I used a cnn to perform the experiments because it cross-validated the best on both datasets.

3.The results of the model predictions reveal some interesting trends, such as sad songs reaching a 10-year high in 2020 and 2022 and gradually decreasing from happy songs in 2020 to a near 5-year low in 2022, possibly reflecting global negative emotional events of the COVID-19 pandemic and the Russia-Ukraine conflict in 2022.

Evaluation and Discussion: (PPT21)

1-5 The project involved meticulous dataset preparation, including data acquisition, cleaning, and balancing. Reproducing key research helped me understand the original approach and identify areas for improvement. According to the direction that understanding can be improved, remarkable findings are made in three aspects: experimental embedding techniques, preprocessing techniques, and audio feature integration. The results highlight the importance of choosing an appropriate embedding technique and preprocessing strategy, as well as the value of integrating audio features, especially the generalization of the model when the dataset is diverse.

6.Practical application of the model on Spotify top 100 songs demonstrates its real-world effectiveness and provides insights into global sentiment trends. This application not only verifies the accuracy of the model, but also validates its generality and relevance in real scenarios.

7.Conclusion: My research journey was comprehensive, from dataset preparation to model development and practical application. This project highlights the importance of a systematic approach and continuous optimization in improving music sentiment analysis models. The fusion of lyric and audio features proved to be a key factor in improving the performance and accuracy of the model.

Next Steps: (PPT22)

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

Complete a comprehensive report on the findings, contributing to the field of Music Emotion Recognition.