Term Project Final Report

MIS 49Y Applied Machine Learning
Fall 2023

Emotion Recognition Based on Lyrics

Alara Balaman

Problem Statement and Goal:

Problem Statement:

Emotion recognition in music is a challenging yet essential task with significant applications in various domains, including personalized music recommendation systems, mood-based playlist generation, and understanding the emotional impact of music on listeners. The fusion of lyrics and the associated embeddings within Russell's model of affect (Valence-Arousal-Dominance tags) provides a rich source of information for training machine learning models to predict emotional categories associated with songs. Despite the progress made, achieving high accuracy in emotion recognition remains elusive, and the choice of feature representation and classifier greatly influences the model's performance.

Goal:

The primary goal of our project is to enhance the accuracy and robustness of emotion recognition in songs based on lyrics and the associated embeddings. We intend to further investigate the factors influencing the model's performance, including feature engineering, model selection, and hyperparameter tuning. The identified challenges and opportunities will guide us towards refining our approach and achieving a more reliable and versatile emotion recognition system. We aim to push the boundaries of emotion recognition in songs, providing a valuable tool for music enthusiasts, researchers, and industries relying on emotional content understanding in music. The insights gained from this project will contribute to the broader field of affective computing and pave the way for more accurate and reliable emotion recognition systems in the context of music analysis.

Literature Review

Song lyrics are a critical component in emotion detection, as they convey significant emotional cues. The interpretation of these lyrics involves understanding the

underlying emotional tone, which is often represented in terms of valence (positive or negative emotional charge) and arousal (intensity of the emotion). These variables are essential for classifying the emotional content of songs. (Yang, 2021).

Emotion detection in song lyrics is grounded in theories of linguistic analysis and emotional expression in music. This involves understanding the emotional tone conveyed by the combination of words, phrases, and overall narrative of the lyrics. This area intersects with aspects of psychology, where the emotional impact of language is studied, and computational linguistics, where these principles are applied to develop algorithms capable of detecting emotions (Song et al., 2012)

The key concepts in this domain include sentiment analysis, natural language processing (NLP), and emotional categorization in music. Sentiment analysis involves assessing the affective nature of language, while NLP applies machine learning and computational linguistics to process and interpret human language. (Mohammad, 2021) Emotional categorization in music refers to the classification of songs into different emotional categories based on their lyrical content.

Revathy et al. (2023) tackle challenges in music emotion recognition, carrying out The LyEmoBERT's experiment using BERT-Base, and overfitting was prevented by adding a dropout layer. The average accuracy achieved by the LyEmoBERT model by running the experiment for 10 times using the above-mentioned batch sizes and learning rates has resulted as 92%.

Malheiro et al. (2019) focused on Classification and Regression of Music Lyrics where the work was done according to Russell’s Emotion Model and achieved 73.6% accuracy

Prediction of Genres and Emotions by Song Lyrics by Stanford ICME students was carried out using fine-tuned pre-trained BERT model and applied transfer learning with 4 emotion categories as output, which gave a maximum accuracy of 32.5%, the accuracy being that low was due to BERT requiring a lot of computation power and them choosing a relatively small proportion of their data.

Zhou (2022) achieved a 66.87% F1-score with the SVM model on the AllMusic Dataset with lowercase
conversion, noise removal, stop-words removal, and stemming. The BERT model obtained 63.59%, and the XLNet model achieved 70.09%. Their experiments showed transfer learning methods do not always work better than traditional machine learning methods when the dataset is relatively small.

Rachman et al. (2019) developed a rule-based method for detecting song emotion using arousal and valence values derived from audio and lyrics features. Their analysis reveals that audio features represent valence better, while lyrics features excel in capturing arousal, providing insights into the complementary roles of these modalities.

Padmane et al. (2022) explore mood categorization of songs based solely on lyrics using Decision Tree and Random Forest models though their small dataset size limited the results.

Reddy et al. (2018) introduce code-mixing features to enhance sentiment prediction in song lyrics leveraging a language identification tool and observed 4-5% improved accuracy compared to traditional approaches.

Raschka, S. (2016) presents a sentiment prediction system using a naive Bayes classifier based on song lyrics. The study focuses on detecting a happy mood with high precision, showcasing the potential applications of sentiment prediction in contemporary music.

Dataset Description

MuSe Musical Sentiment Dataset:

The MuSe dataset comprises sentiment information for 90,001 songs, deriving sentiment from social tags on Last.fm, informed by the Warriner et al. database. Sentiments are expressed across three dimensions: valence, arousal, and dominance. Last.fm tags help infer song genres by comparison to a predefined list. The dataset serves as a proof-of-concept, with a work-in-progress disclaimer. Duplicates may arise due to tag diversity issues in data collection. The dataset creation process is detailed in the paper by Akiki and Burghardt (2020)

Genius Song Lyrics Dataset:

The Genius dataset, scraped as recent as 2022, builds upon the 5 Million Song Lyrics Dataset. It provides information from Genius, a platform for song annotations. The dataset includes song titles, genres, artists, release years, page views, features, and lyrics. The lyrics may require preprocessing due to metadata present within the text. Language information is provided, with columns indicating the language according to CLD3 and FastText's langid. The dataset serves well for NLP tasks and offers opportunities for data cleaning practice. The dataset consists of various types of pieces, and the language column combines information from CLD3 and FastText with non-NaN entries only when both methods agree.

Combined Dataset:

A new dataset was created by merging lyrics from the Genius dataset with the MuSe dataset, resulting in 29,130 songs. Null values were cleaned, and lyrics were preprocessed, three language columns are scaled down into one language column which already was the combination of the other two. Emotion labels for classification were derived from the "seeds" column, which contains 262 emotions. Six broad emotion categories were created and encoded. The final dataset includes columns such as track, artist, seeds, number_of_emotion_tags, valence_tags, arousal_tags, dominance_tags, genre, tag, broad_category, cleaned_lyrics, and broad_category_encoded. The resulting dataset is shaped (29130, 12), combining sentiment information and lyrics, providing a comprehensive resource for lyrical sentiment analysis with machine learning.

Methodology

Cleaning and Preparing the Song Lyrics Data

The 'seeds' column was transformed into actual lists, and a process was undertaken to extract unique emotions from this column, which comprised 262 distinct emotion representation labels. Subsequently, a mapping function was employed to categorize specific emotions into six broader categories.

Handling Class Imbalance

During the initial phase, an approach was adopted where records matching more than one category were assigned a hierarchy among the categories. This approach, however, led to a disproportionate representation of records in one category (Energetic/Vibrant) compared to the others, resulting in imbalanced classes. To address this issue, a decision was made to omit records associated with multiple categories. This refinement led to the establishment of five balanced, unique broader categories, namely Happy/Fun, Sad/Anxious, Romantic/Loving, Calm/Reflective, and Aggressive/Enigmatic. This approach ensured equitable representation and balanced training across diverse emotional states.

Text Vectorization

The dataset underwent a preprocessing phase where the lyrics were standardized. This process entailed converting the text to lowercase, eliminating stopwords, and expunging special characters along with generic textual elements such as [intro], [chorus], [bridge], etc. Following this cleansing procedure, the lyrics column was subjected to vectorization using Term Frequency-Inverse Document Frequency (TF-IDF) coupled with n-grams, specifically setting the n-gram range to (1,2). This methodological approach enabled the transformation of the textual data into a format conducive for further computational analysis.This vectorized text data was combined with other numeric features. The dataset was then split into 80% training and 20% test examples.

Model Development

We commenced our modeling by training the cleaned lyrics. Multiple machine learning algorithms were evaluated, including Random Forest Classifier, XGB Classifier, SVM, Randomized Search CV, Gradient Boosting Classifier, Multinomial Naive Bayes, and K-Nearest Neighbors. The Random Forest Classifier emerged as the most successful, demonstrating superior performance compared to other models.

Incorporation of Embeddings

Following the successful initial model, we extended our analysis by incorporating three embeddings - 'valence_tags', 'arousal_tags', and 'dominance_tags' - into the feature set alongside 'cleaned_lyrics'. The combination aimed to enhance the model's predictive capabilities by considering emotional dimensions within Russell's model of affect.

Hyperparameter Tuning

To enhance classification performance, we employed a RandomForestClassifier and conducted hyperparameter tuning using a grid search. The hyperparameter grid included variations in 'n_estimators,' 'max_depth,' 'min_samples_split,' and 'min_samples_leaf.' After thorough experimentation, the optimal hyperparameter values were determined as follows: 'max_depth': None, 'min_samples_leaf': 1, 'min_samples_split': 5, 'n_estimators': 200.

Experiment Results

Initial attempts at classification using classifiers such as XGBoost and RandomForest yielded an accuracy of approximately 30% when considering only the preprocessed and cleaned lyrics column. After addressing the imbalance caused by the Energetic/Vibrant category, we started getting much more improved results.

In the conducted experiments, several machine learning models were applied to the task of classifying the dataset, utilizing different combinations of features and methodologies. The focus was on assessing the efficacy of each model in terms of its accuracy.

Multinomial Naive Bayes

Initially, a Multinomial Naive-Bayes classifier was trained exclusively on the cleaned lyrics. The performance of this model was modest, achieving an accuracy of 0.45. This baseline result provided a preliminary understanding of the dataset's complexity and the challenges in classification based solely on textual data.

XGB Classifier

Subsequently, an XGB Classifier was employed, again trained solely on the cleaned lyrics. This approach yielded a more promising result, with an accuracy of 0.53 and a support of 3840 instances. The improvement indicated the potential of more sophisticated models in handling the task.

Hyperparameter Tuning for XGB Classifier

To optimize the XGB Classifier, hyperparameter tuning was conducted using RandomizedSearchCV. However, this strategy resulted in a slight decrease in

accuracy, recording a figure of 0.47 with the same support of 3840. This outcome suggested that the default parameters of the XGB Classifier were already near-optimal for this specific dataset.

Support Vector Machine (SVM)

Training an SVM with the same combination of features (valence, arousal, dominance, and cleaned lyrics) was also explored. This model resulted with 0.48 accuracy.

Enhanced XGB Classifier

The incorporation of additional features such as valence, arousal, dominance, along with the cleaned lyrics, significantly improved the results for the XGB Classifier. Two trials were conducted, yielding accuracies of 0.67 and 0.64, respectively. This enhancement underscored the importance of incorporating emotion-related features in conjunction with textual data for improved classification performance.

Model Improvement through Feature Inclusion

Initially, the model was trained solely with the cleaned lyrics column, achieving an accuracy of 66%. To explore the impact of additional features, we extended the model to include 'number_of_emotion_tags,' 'valence_tags,' 'arousal_tags,' and 'dominance_tags' along with the cleaned lyrics. This extension significantly improved accuracy, raising it to 70%.

RandomForest Classifier

The application of a RandomForest Classifier, utilizing the combination of valence, arousal, dominance values, and cleaned lyrics, marked a significant improvement, achieving an accuracy of 0.70. This model demonstrated the effectiveness of ensemble learning techniques in managing the dataset's complexity.

Best Performing Model

The most effective model was the RandomForest Classifier, which, when trained with valence, arousal, dominance, and cleaned lyrics, achieved the highest accuracy of 0.79. This result represents a substantial advancement from the initial models and underscores the RandomForest Classifier's robustness and suitability for this classification task.

Acknowledging the persisting imbalance in the Energetic/Vibrant category, we redistributed examples from this category to the other five emotion categories. Instances with multiple emotion tags were discarded, resulting in a final dataset of

19,801 examples. The RandomForestClassifier was then retrained using the best hyperparameter values, achieving an accuracy of approximately 66%.

Further, we expanded the model to incorporate 'valence_tags,' 'arousal_tags,' 'dominance_tags,' and 'cleaned_lyrics.' This comprehensive model yielded outstanding results, reaching an impressive accuracy of 78.5%. This performance was notably higher than previous models and positioned our classifier as one of the most effective models reported in the literature.

Evaluation Metrics and Confusion Matrix

The effectiveness of our final model was evaluated using various metrics, including precision, recall, and F1-score, calculated for each emotion category. For the model trained solely with cleaned lyrics, the results were as follows:
- Accuracy: 66.47%
- Precision: 67% (macro average)
- Recall: 67% (macro average)
- F1-score: 66% (macro average)

The confusion matrix further emphasized the model's proficiency, with a discernible dark blue diagonal indicating accurate predictions and white squares surrounding the diagonal illustrating minimal misclassifications.

For the extended model incorporating 'valence_tags,' 'arousal_tags,' 'dominance_tags,' and 'cleaned_lyrics,' the evaluation metrics were even more impressive:
- Accuracy: 78.5%
- Precision: 79% (macro average)
- Recall: 79% (macro average)
- F1-score: 79% (macro average)

The confusion matrix reinforced the model's robustness, with accurate predictions across all emotion categories and minimal instances of misclassification.

In conclusion, our study demonstrates the effectiveness of utilizing lyrics and additional emotion-related features for predicting musical sentiment. The refined RandomForestClassifier, trained with optimal hyperparameters and an augmented feature set, produced exceptional results, outperforming existing models in the

literature. This model holds promise for applications in music recommendation systems and emotional analysis of large-scale music databases.

Conclusion:

RandomForestClassifier, coupled with hyperparameter tuning through a grid search, resulted in an optimized model with notable accuracy. Moreover, the inclusion of additional features such as 'number_of_emotion_tags,' 'valence_tags,' 'arousal_tags,' and 'dominance_tags' alongside the cleaned lyrics significantly improved classification performance.

Recognizing the persisting imbalance, we further refined our dataset and retrained the RandomForestClassifier, achieving an impressive accuracy of 78.5%. This final model not only outperformed previous iterations but also positioned itself as a standout in the literature. The incorporation of 'valence_tags,' 'arousal_tags,' 'dominance_tags,' and 'cleaned_lyrics' contributed to its exceptional precision, recall, and F1-score across all emotion categories.

The thorough evaluation of our model, including precision, recall, and the confusion matrix, underscored its proficiency in accurate emotion prediction. These results offer promising implications for applications in personalized music recommendation systems, mood-based playlist generation, and emotional analysis of extensive music databases.

Discussion and Future Work:

Our project embarked on the ambitious endeavor of emotion recognition in songs, an area that combines the complexities of both linguistic and musical sentiment analysis. By integrating the MuSe musical sentiment dataset with Genius lyrics, our team aimed to build a robust classification model that could effectively identify emotional content in songs. The inherent challenge in this task lies in the nuanced interplay between the lyrics and the musical elements, each contributing significantly to the conveyed emotion.

Our methodology involved the use of TF-IDF for text representation, which, while effective, presented limitations in capturing the semantic richness of lyrics. We complemented this approach by incorporating arousal-valence-dominance (AVD) tags, which provided an additional dimension to the emotional context. This combination allowed for a more comprehensive representation of the song's emotional content.

Furthermore, we employed hyperparameter tuning on a Random Forest classifier, a decision that proved instrumental in optimizing the model's performance. The results, while promising, suggested that there is room for improvement in the model's ability to discern complex emotional states.

Looking ahead, we propose the adoption of BERT (Bidirectional Encoder Representations from Transformers) for text cleaning and vectorization in future iterations of the model. The decision to exclude BERT in this instance was driven by its computational intensity. However, we hypothesize that BERT's state-of-the-art capabilities in understanding contextual nuances in text could significantly enhance the model's accuracy. This integration could potentially revolutionize the field by providing a more sophisticated understanding of the interplay between lyrics and emotion.

References:

- Akiki, C., & Burghardt, M. (2020). Toward a Musical Sentiment (MuSe) Dataset for Affective Distant Hearing. *CHR 2020*, 225–235. http://ceur-ws.org/Vol-2723/short26.pdf
- Yang, J. (2021). A novel music emotion recognition model using neural network technology. Frontiers in Psychology, 12. https://doi.org/10.3389/fpsyg.2021.760060
- Song, Y., Dixon, S., & Pearce, M. T. (2012). EVALUATION OF MUSICAL FEATURES FOR EMOTION CLASSIFICATION. ISMIR 2012, 523–528. http://ismir2012.ismir.net/event/papers/523_ISMIR_2012.pdf
- Mohammad, S. M. (2021). Sentiment analysis. In Elsevier eBooks (pp. 323–379). https://doi.org/10.1016/b978-0-12-821124-3.00011-9

- Revathy, V. R., Pillai, A. S., & Daneshfar, F. (2023). LyEmoBERT: Classification of lyrics' emotion and recommendation using a pre-trained model. Procedia Computer Science, 218, 1196–1208. https://doi.org/10.1016/j.procs.2023.01.098
- Li, S. (2021). Prediction of Genres and Emotions by Song Lyrics.
- Zhou, Y. (2022). Music Emotion Recognition on Lyrics Using Natural Language Processing. McGill University (Canada).
- Mohammad, S. M. (2021). Sentiment analysis. In Elsevier eBooks (pp. 323–379). https://doi.org/10.1016/b978-0-12-821124-3.00011-9
- R. Malheiro, R. Panda, P. Gomes and R. P. Paiva, "Emotionally-relevant features for classification and regression of music lyrics", IEEE Transactions on Affective Computing, vol. 9, no. 2, pp. 240-254, 2019.
- Sharma, H., Gupta, S., Sharma, Y., & Purwar, A. (2020, March). A new model for emotion prediction in music. In 2020 6th International Conference on Signal Processing and Communication (ICSC) (pp. 156-161). IEEE.
- Rachman, F. H., Samo, R., & Fatichah, C. (2019, October). Song emotion detection based on arousal-valence from audio and lyrics using rule based method. In 2019 3rd International Conference on Informatics and Computational Sciences (ICICoS) (pp. 1-5). IEEE.
- Padmane, P., Agrahari, K., Kesharwani, R., Mohitkar, K., Agrahari, M. K., Khan, S., & Kamale, N. (2022). IMPLEMENTATION OF PREDICTION OF SONG MOOD THROUGH LYRICS. Open Access Repository, 9(6), 137-141.
- Reddy, G. R. R., & Mamidi, R. (2018). Addition of code mixed features to enhance the sentiment prediction of song lyrics. arXiv preprint arXiv:1806.03821.
- Raschka, S. (2016). MusicMood: Predicting the mood of music from song lyrics using machine learning. arXiv preprint arXiv:1611.00138.