Multi-Emotion Classification of Song Lyrics

Tin Popović

Tin Popović, University of Zagreb – FER, Zagreb Croatia

Abstract

Purpose – This research aims to explore the efficacy of multi-label emotion classification in song lyrics, focusing specifically on the genres of R&B and Hip-Hop. The study seeks to understand the emotional spectrum expressed in these genres and how advanced machine learning algorithms can accurately classify these emotions.

Design/Methodology/Approach – The study employs a comparative analysis of three machine learning models: Naive Bayes, Random Forest, and a fine-tuned version of RoBERTa (a robust transformer-based model). The models are trained and tested on a dataset comprising a diverse range of R&B and Hip-Hop song lyrics. The classification task involves multiple emotions, recognizing the complexity and nuance in lyrical content. The performance of each model is evaluated based on accuracy, precision, recall, and F1 score.

Findings – The results demonstrate that the fine-tuned RoBERTa model outperforms Naive Bayes and Random Forest in terms of F1 score and precision, indicating its superior capability in capturing the subtleties of emotional expressions in song lyrics. The study also reveals interesting patterns and emotional profiles unique to R&B and Hip-Hop genres, underscoring the rich emotional depth in these musical styles.

Originality/Value – This research contributes to the field of music information retrieval and computational musicology by providing insights into emotion classification in song lyrics, a relatively underexplored area. The findings highlight the potential of advanced machine learning techniques, especially transformer-based models like RoBERTa, in understanding the emotional dimensions of music. This study offers valuable implications for music recommendation systems, sentiment analysis in lyrics, and enhances our understanding of the emotional impact of R&B and Hip-Hop music.

Keywords – Machine Learning; Deep Learning; Song Lyrics; Music; Data; F1; Precision; Recall

Paper Type – Research paper.

1. Introduction

The interaction between emotion and music forms a fundamental aspect of human cultural expression, where song lyrics frequently act as a powerful medium for conveying a diverse array of emotional experiences. This is particularly pronounced in genres such as R&B and Hip-Hop, which not only offer rich emotional narratives but also often reflect the social and cultural dynamics of their times. The perception of Hip-Hop has been marked by a contentious discourse, often stereotyped with themes of violence and aggression. This research aims to delve deeper into these genres, employing advanced machine learning techniques - Naive Bayes, Random Forest, and a fine-tuned RoBERTa model - to perform a multi-label emotion classification of song lyrics. The objective is to quantitatively analyze the emotional content within these genres, challenging or substantiating common perceptions with data-driven insights. By focusing specifically on R&B and Hip-Hop, this study seeks to unravel the complex tapestry of emotions that go beyond the often assumed narrative of violence in the Hip-Hop community. This exploration not only contributes to the fields of computational musicology and sentiment analysis but also offers a nuanced understanding of how different machine learning models capture and interpret the emotional undertones in music lyrics. Thus, this

research stands at a critical intersection of technology, music, and cultural studies, aiming to shed light on the emotional landscape of these influential music genres and to contribute to a more nuanced understanding of their societal impact.

2. Related work

In the realm of multi-label emotion classification in song lyrics, recent research has advanced various methodologies, reflecting the complexity of this task.

One pioneering study by Mihalcea and Strapparava (Mihalcea, 2010) utilized basic text classification methods to predict emotions in lyrics, setting a foundational approach for subsequent research. A study by (Darren Edmonds, 2021) emphasizes the advantages of using specialized, smaller datasets for song emotion classification over larger, generalized datasets, suggesting a targeted approach for improved performance. Complementing this, research from the (Konstantinos Trohidis, 2011) on Audio, Speech, and Music Processing explores innovative ranking-based methods such as Ranking by Pairwise Comparison (RPC) and Calibrated Label Ranking (CLR). These methods address the challenges of classifying multiple emotions in music, showcasing the need for nuanced approaches in emotion detection.

A notable study (Jia, 2022) focuses on a CNN-LSTM based model for lyrics emotion classification, underlining the effectiveness of BiLSTM structures in enhancing classification accuracy by leveraging bidirectional data analysis. Furthermore, ResearchGate highlights a multi-modal approach that integrates audio and lyrical analysis for music emotion classification, emphasizing the importance of combining various data forms to enrich emotional expression analysis in music.

(Tsaptsinos, 2017) focused on using a Hierarchical Attention Network (HAN) for lyrics-based music genre classification, comparing different models such as Majority Classifier, Logistic Regression, Long Short-Term Memory (LSTM), and Hierarchical Networks. The research found that neural-based models like LSTM and HAN generally outperformed simpler models. The HAN model, especially when applied at the line level, showed better results compared to segment-level analysis, indicating the effectiveness of a more detailed lyrical analysis. This study highlights the significance of model complexity and its relation to classification accuracy in the context of music genre classification using lyrics.

Another significant study in the field of multi-label classification of song lyrics is by (Mayer N. , 2008), who explored the combination of audio and lyrics features for genre classification in digital audio collections. This research was an early attempt to integrate different modalities, such as lyrics and audio features, to improve the accuracy of music genre classification. The study underscores the potential of combining lyrical content with audio analysis to provide a more holistic understanding of musical genres.

Similarly, (Mayer R. , 2011) extended this approach by proposing musical genre classification using ensembles of audio and lyrics features. Their work focused on leveraging the complementary nature of the lyrical and audio aspects of music to enhance the genre classification process. This approach reflects the growing trend in the field to utilize multi-modal data for more accurate and comprehensive music analysis.

Furthermore, the adoption of deep learning techniques for this task has seen widespread use, employing models ranging from Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) network (Rémi Delbouys, 2018) (Jiddy Abdillah, 2020) to advanced

transformer-based architectures like BERT and ELMo (Loreto Parisi, 2019) (Gaojun Liu, 2020) Liu and Tan, 2020).

A multimodal strategy, integrating both audio and lyrical data, has also been pursued for emotion detection in music. (Rada Mihalcea, 2012) introduced a corpus that includes both music and lyrics, achieving encouraging results by utilizing representations from both modalities for emotion classification.

However, reliance on audio data has been criticized for potentially introducing bias in emotion classification. Studies by Susino and Schubert (Schubert, 2019) have shown that genres like heavy metal and hip-hop are often stereotypically associated with more negative emotions compared to pop music, despite having similar lyrical content. This bias was investigated by Fried (Fried., 1999) and Dunbar (Adam Dunbar, 2016) who found that identical lyrics were perceived differently based on the genre, with rap music being viewed as more offensive compared to country music.

The critical role of lyrics in accurately predicting emotions in music has been underscored by several studies. Yang and Lee (Lee, 2009) successfully transformed song lyrics into psychological feature vectors, demonstrating that lyrics alone could create effective and interpretable classification models. Hu (Xiao Hu) and later Hu and Downie (Downie, 2010) , found that lyrical features often surpassed audio features in mood prediction accuracy across several mood categories, challenging the assumption that combining lyrical and audio features necessarily leads to better mood prediction outcomes.

In sum, these studies collectively contribute to the field by offering diverse perspectives on data utilization, methodological approaches, and advanced computational techniques. They underscore the need for specificity in dataset selection, integration of multimodal data, and innovative classification methods, all essential for the nuanced task of emotion classification in song lyrics.

3. Methodology

3.1 Dataset description

The dataset for the "MultiLabel Emotion Classification on Song Lyrics" project was meticulously assembled by combining two distinct sources to create a comprehensive and diverse collection of song lyrics. The first part of the dataset originates from the Edmonds Dance dataset, a pre-processed and well-curated collection of song lyrics specifically designed for multi-emotion classification, as detailed in (Darren Edmonds, 2021) From this dataset, a subset of 31 songs was carefully selected. The second part of the dataset was independently compiled by scraping the GENIUS (Genuis, 2009) website to gather song lyrics from a targeted selection of artists, focusing predominantly on the RnB and Hip-Hop genres. This effort resulted in the addition of 203 songs, enriching the dataset with contemporary and genre-specific content. The process of data acquisition from the GENIUS website to compile a repository of Hip-hop and R&B songs adhered to a systematic methodology. Initially, an API token was procured through official channels, thereby establishing authorized access to GENIUS resources. Subsequently, leveraging the "*lyricsgenius*" library within the Python environment, the Genius class was instantiated, enabling the invocation of the search_artist function. This function facilitated the retrieval of pertinent song data, contingent upon specified parameters such as the artist's name and the desired number of songs, with optional sorting

criteria based on popularity metrics. Upon retrieval, the acquired data, encompassing song IDs, titles, artists, and lyrical content, was meticulously cataloged into distinct lists. Finally, collected songs and related information were saved for later usage and processing.
The collected lyrics were initially labeled with emotions using OpenAI's GPT-3 Turbo API, an advanced natural language processing tool, incurring a modest cost of approximately $7.25. This automated labeling was subsequently refined and verified manually to ensure accuracy and relevance.
The final dataset is structured into columns representing *song_id, artist, title, lyric*, and a spectrum of emotions including *anger, confidence, desire, disgust, gratitude, joy, love, lust, sadness, shame, fear, and anticipation*. This dual-source approach not only enhances the dataset's diversity but also ensures a blend of automated efficiency and meticulous human oversight in the emotion labeling process.

	song_id	artist	title	lyric	anger	confidence	desire	disgust	gratitude	joy	love	lust	pride	sadness	shame	fear	surprise	anticipation
0	3315890.0	Drake	God’s Plan	and they wishin' and wishin' and wishin' and w...	0	1	0.0	0	1.0	1	0.0	0.0	0.0	0	0.0	0.0	0.0	0.0
1	3807759.0	Drake	In My Feelings	trap, trapmoneybenny this shit got me in my fe...	0	0	0.0	0	0.0	1	1.0	1.0	0.0	0	0.0	0.0	0.0	0.0
2	2263723.0	Drake	Hotline Bling	you used to call me on my you used to, you use...	0	0	1.0	0	0.0	0	0.0	0.0	0.0	1	0.0	0.0	0.0	0.0
3	2450584.0	Drake	One Dance	baby, i like your style grips on your waist,...	0	0	1.0	0	0.0	1	0.0	0.0	0.0	0	0.0	0.0	0.0	0.0
4	200546.0	Drake	Hold On, We’re Going Home	i got my eyes on you you're everything that i ...	0	0	1.0	0	0.0	0	1.0	0.0	0.0	0	0.0	0.0	0.0	0.0

Figure 1 Example of Dataset records

3.2 Dataset Preprocessing

The preprocessing of the dataset for the "MultiLabel Emotion Classification on Song Lyrics" project involved a series of meticulous steps to ensure data quality and consistency. The Edmonds Dance dataset, which formed the first part of the data, required no additional text preprocessing due to its already clean state. However, the lyrics collected via web-scraping presented a different challenge. These lyrics underwent a thorough cleaning process to enhance their usability for emotion classification. This process involved converting all lyrics to lowercase and removing any non-lyrical content such as words within brackets (often indicating song parts like intro or chorus), as well as any metadata like contributor names. Additionally,

new lines were replaced with spaces, and any trailing numbers or special characters were removed to ensure a uniform text format across the dataset.

Following the cleaning of the lyrics, a One-Hot Encoding (OHE) operation was performed on the emotion columns to facilitate better analysis and model training. This step transformed the categorical emotion labels into a binary matrix representation, making it suitable for the multi-label classification task. The two datasets were then combined, resulting in a comprehensive collection of songs spanning various emotions and genres.

Upon combining the datasets, it was noted that there were no NaN values present, indicating a complete and robust dataset. An analysis of the occurrence of emotions in the dataset revealed that the emotions of pride and surprise were significantly underrepresented, occurring only 2 and 1 times, respectively. Consequently, these emotions were dropped from the dataset. However, as the songs associated with these emotions also contained other emotion labels, their exclusion did not result in the loss of any songs.

To address the issue of imbalance in emotion representation, an oversampling strategy was employed. This was particularly crucial as the project dealt with a multi-label problem, limiting the applicability of certain techniques like SMOTE. The goal was to ensure that each emotion had at least 20 examples, thus balancing the dataset and improving the model's ability to learn from underrepresented emotions. As a result of these preprocessing steps, the final dataset encompassed 289 songs with the following distribution of emotions: anger (60), confidence (117), desire (62), disgust (48), gratitude (20), joy (68), love (51), lust (36), sadness (83), shame (20), fear (20), and anticipation (20). This balanced and well-processed dataset provides a solid foundation for accurate and effective emotion classification in song lyrics.

EMOTION	FREQUENCY
Anger	60
Confidence	117
Desire	62
Disgust	48
Gratitude	20
Joy	68
Love	51
Lust	36
Sadness	83
Shame	20
Fear	20
Anticipation	20

Table 1. Emotion frequencies in the collected dataset

3.3Theoretical backround

The theoretical foundation of the "MultiLabel Emotion Classification on Song Lyrics" research incorporates a blend of classical and contemporary machine learning algorithms to effectively classify emotions in song lyrics. The study employs three distinct methods: Naive Bayes, Random Forest, and fine-tuning of RoBERTa (a robustly optimized BERT approach).

3.3.1 Naïve bayes

This study utilizes a Naive Bayes model within a Classifier Chain framework, integrated through a pipeline approach for emotion detection in song lyrics. The theoretical foundation of this approach is rooted in Bayesian statistics, where the Naive Bayes classifier applies the Bayes theorem with the "naive" assumption of independence between every pair of features. This model is particularly suited for text classification due to its simplicity, efficiency, and the ability to handle high-dimensional data.

$$P(y_k|x) = \frac{P(x|y_k) * P(y_k)}{P(x)}$$

The Multinomial Naive Bayes variant is employed, which is tailored for feature vectors representing the frequencies with which events have been generated by a multinomial distribution. This is appropriate for text data, where features are typically the counts or term frequencies of words or n-grams in the document.
A critical component in preprocessing text for such models is the transformation of raw text into a more digestible form for the algorithm. This is achieved using the Term Frequency-Inverse Document Frequency (TF-IDF) Vectorizer. The Term Frequency-Inverse Document Frequency (TF-IDF) Vectorizer is a pivotal technique in the field of text mining and information retrieval, designed to reflect the importance of a word within a document in the context of a collection of documents or corpus. The principle behind TF-IDF comprises two components: Term Frequency (TF) and Inverse Document Frequency (IDF). Term Frequency measures the frequency of a word within a single document, signifying the importance of the word relative to the document's length. The assumption here is that the more frequently a term appears in a document, the more significant it is for that document. However, to balance the scale and reduce the weight of terms that appear too commonly across all documents (thereby offering little unique information), the Inverse Document Frequency is introduced. IDF diminishes the weight of terms that are common across the corpus and increases the weight of terms that are rare. The TF-IDF value is a product of these two statistics, ensuring that words are properly weighted according to both their local relevance and their rarity across documents. This dual consideration makes TF-IDF an effective method for converting text into a numerical representation that highlights the most distinguishing words for each document within the larger context of the document set, facilitating nuanced text analysis and machine learning models' understanding of language.

$$TF - IDF(t, d, D) = TF(t, d) \times IDF(t, D)$$

$$TF(t, d) = \frac{Total\ number\ of\ terms\ in\ document\ d}{Frequency\ of\ term\ t\ in\ document\ d}$$

$$IDF(t, D) = log(\frac{Number\ of\ documents\ containing\ term\ t}{Total\ number\ of\ documents\ in\ set\ D})$$

The model's hyperparameters, such as the maximum features for the TF-IDF Vectorizer and the smoothing parameter (alpha) for the Naive Bayes classifier, are meticulously optimized through Grid Search with Cross-Validation. The grid search explores a range of values to find the best combination of parameters that yield the highest F1 macro score, a balanced metric for evaluating models on imbalanced datasets. The alpha parameter is crucial for smoothing, preventing zero probabilities in further calculations and allowing the model to handle unseen words during training.

This comprehensive approach, combining TF-IDF for feature extraction, Multinomial Naive Bayes for classification, demonstrates a robust methodology for emotion detection in song lyrics. Through the meticulous tuning of hyperparameters and the application of a theoretically grounded model, this study contributes valuable insights into the emotional landscapes embedded within lyrical compositions.

3.3.1 Random Forest

The Random Forest algorithm is a powerful ensemble learning method used for both classification and regression tasks, which operates by constructing a multitude of decision trees at training time and outputting the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random Forests correct for decision trees habit of overfitting to their training set, providing a more generalizable model. In the context of emotion detection in song lyrics, the Random Forest classifier leverages the strength of multiple decision trees to improve prediction accuracy and robustness. The ensemble method works by creating many decision trees from randomly selected subsets of the training set and features. Each tree in the forest is built from a sample drawn with replacement (bootstrap sample) from the training set. Furthermore, when splitting a node during the construction of the tree, the split is chosen from a random subset of the features, rather than the best split among all features. This randomness helps to make the model more diverse, reducing the variance without increasing the bias, which in turn lowers the risk of overfitting.
The theoretical foundation of Random Forests also involves aggregating the predictions of individual trees through "bagging" (Bootstrap Aggregating) to reduce variance. Each tree votes for a class, and the class receiving the most votes becomes the model's prediction for ensemble methods in classification problems, described mathematically as

$$RF_{Class(x)} = Mode\ \{Tree_1(x), Tree_2(x), \dots, Tree_N(x)\}.$$

Key hyperparameters in Random Forest include the number of trees in the forest, the maximum depth of the trees, and the maximum number of features considered for splitting at each leaf node. Tuning these hyperparameters is crucial for optimizing the model's performance. A higher number of trees can increase the model's accuracy up to a point but also increases computational complexity. The maximum depth of the trees controls the complexity of the model, with deeper trees potentially capturing more intricate patterns but also increasing the risk of overfitting. The choice of maximum features affects the diversity of the trees, with too low a value causing the trees to be too similar.
In the applied implementation, the TF-IDF (Term Frequency-Inverse Document Frequency) Vectorizer is utilized for preprocessing text data, converting the raw song lyrics into a matrix of TF-IDF features. This transformation emphasizes important words and allows the Random Forest model to work with numerical data, enhancing its ability to classify the emotional content of the lyrics effectively.
The combination of TF-IDF for feature extraction and Random Forest for classification, refined through Grid Search with Cross-Validation for hyperparameter optimization, embodies a robust approach to emotion detection in song lyrics. This methodology not only capitalizes on Random Forest's ability to handle high-dimensional data and its resilience to overfitting but also leverages the nuanced representation of text data provided by TF-IDF, offering a theoretically and empirically sound framework for analyzing emotional expressions in text.

3.3.1 RoBERTa

The RoBERTa (Robustly Optimized BERT Approach) model represents an advanced iteration of the BERT (Bidirectional Encoder Representations from Transformers) architecture, specifically designed to improve on BERT's language understanding capabilities. Leveraging a transformer-based mechanism, RoBERTa optimizes the pre-training process through modifications such as dynamic masking, larger mini-batches, and longer sequences, which collectively enhance its performance on downstream tasks, including multi-label classification. In the context of emotion detection in song lyrics, RoBERTa's ability to understand the context of each word within an entire sequence of lyrics makes it particularly suited for capturing the nuanced emotional landscape conveyed in textual data.

For the specific task of classifying multiple emotions in song lyrics, the model was adapted from a pre-trained RoBERTa base specifically fine-tuned for emotion detection (Lowe, Hugging Face, 2022)indicating that it has been previously optimized on a related task, thereby leveraging transfer learning to improve accuracy on the target dataset. The model was configured to predict 12 distinct emotions, aligning with the number of labels in the dataset, an essential step for tailoring the model to the multi-label classification framework.

Key hyperparameters and configurations for this implementation include the number of training epochs (*num_train_epochs*=7), indicating the total number of times the training dataset is passed through the model, and the batch sizes for training (*per_device_train_batch_size*=6) and evaluation (*per_device_eval_batch_size*=2), which determine the number of samples processed before the model's internal parameters are updated. The choice of these hyperparameters significantly impacts the model's learning efficiency and its ability to generalize from the training data without overfitting.

In this implementation for emotion detection in song lyrics, the AutoTokenizer from the transformers library is utilized (Hugging Face, n.d.), specifically leveraging the SamLowe/roberta-base-go_emotions pre-trained model. This tokenizer is intricately linked to a RoBERTa model that has been fine-tuned on the *GoEmotions* dataset (Lowe, Hugging Face, n.d.), which is a comprehensive dataset curated for the purpose of emotion analysis in text. The choice of this tokenizer is strategic, as it is tailored to the nuances of emotional expression in text, making it exceptionally suited for tasks involving emotion detection.

The custom Song_Dataset class facilitates the preprocessing of lyrics into a format compatible with RoBERTa, including tokenization and encoding with attention masks, ensuring that the model can efficiently process variable-length texts. The use of a sigmoid function in the multi_labels_metrics function allows for the conversion of the model's output logits into probabilities, which are then thresholded to determine the presence of each emotion within the lyrics, providing a nuanced assessment of the model's predictive performance across the spectrum of emotions.

The evaluation of the model's performance on the validation dataset is encapsulated through custom metrics, including hamming loss, precision, recall, and F1 score for each emotion label. These metrics offer a comprehensive view of the model's effectiveness, highlighting its strengths and areas for improvement in distinguishing between the intricate emotional nuances present in song lyrics.

In summary, the adaptation and fine-tuning of the RoBERTa model for multi-label emotion classification in song lyrics exemplify the cutting-edge application of transformer-based models in NLP. Through careful hyperparameter tuning and the strategic use of transfer learning, this approach harnesses RoBERTa's deep contextual understanding, providing a theoretically and empirically robust framework for analyzing emotional expressions in text.

4. Results

In this research, we explored the efficacy of various machine learning models for multi-label classification of song lyrics into emotional categories. Our analysis encompassed three distinct models: Naive Bayes, Random Forest, and RoBERTa fine-tuning.

The Naive Bayes model demonstrated a high precision in classifying certain emotions like 'anger', 'disgust', and 'shame', yet its recall was notably low for most categories, indicating a tendency towards overfitting on specific labels while failing to generalize. This was particularly evident in its poor recall for 'anger' and 'love'. The Random Forest model exhibited a more balanced performance with notable successes in 'confidence' and 'sadness' classification, albeit with shortcomings in detecting 'desire' and 'lust', where it failed to identify any true positives.

Most remarkably, the RoBERTa model, leveraging deep learning and contextual embeddings, outperformed the other models in several aspects. It achieved notably high precision in 'anger', 'joy', 'shame', and 'fear', albeit with varied recall rates, indicating a nuanced understanding of these emotions but with room for improvement in consistently identifying them across different contexts. Notably, RoBERTa struggled in classifying 'gratitude' and 'lust', showing zero precision and recall, which suggests a need for more diverse training data or model fine-tuning for these specific emotions. The balance between precision and recall in RoBERTa was more consistent than in the other models, as evidenced by its overall lower Hamming loss, suggesting a better generalization across multiple labels. However, the presence of certain categories with low recall or F1-scores indicates that while RoBERTa has a strong foundational capability for

this task, it requires further optimization for a more uniform performance across all emotional categories.

	Anger	Joy	Desire	Confidence
Naïve Bayes	1.00	0.62	0.75	0.71
Random Forest	0.86	1.00	0.00	0.73
RoBERT-a	1.00	1.00	0.521	0.75

Table 2 Precision values for top emotions

	Anger	Joy	Desire	Confidence
Naïve Bayes	0.18	0.28	0.14	0.62
Random Forest	0.35	0.22	0.00	0.77
RoBERT-a	0.25	0.052	0.75	0.65

Table 3 Recall values for top emotions.

	Anger	Joy	Desire	Confidence
Naïve Bayes	0.3	0.38	0.14	0.67
Random Forest	0.5	0.36	0.00	0.74
RoBERT-a	0.4	0.1	0.61	0.69

Table 4 F1 scores for top emotions

4.1 Comparison to Darren Edmonds research paper.

In the comparative analysis of multi-emotion classification of song lyrics, this study draws a parallel to the work of Darren Edmonds, with a particular focus on the top four emotions for a more direct and methodologically consistent comparison. Edmonds' research delineates a spectrum of F1 scores across several emotions and models, where Naive Bayes and Lyrics BERT models notably demonstrate proficiency in identifying sadness (0.55 and 0.54, respectively) and joy (0.69 for Lyrics BERT). In contrast, the present study, leveraging both a portion of Edmonds' dataset and newly gathered data, narrows its scope to anger, joy, desire, and confidence. Here, the Random Forest model emerges as a strong contender in detecting anger with an F1 score of 0.5, surpassing the Naive Bayes' performance in Edmonds' study. Conversely, the Naive Bayes model presents a modest decline in joy (0.38) when compared to the robust 0.69 score achieved by Edmonds' Lyrics BERT model. Moreover, the deployment of RoBERTa in this research showcases an exceptional aptitude for discerning desire, with an F1 score of 0.61, in stark contrast to the null results observed for Random Forest in both studies and the CBET BERT in Edmonds' work.

The decision to concentrate on these specific emotions is informed by the methodological parallels and dataset characteristics shared between the two studies, permitting a nuanced evaluation of model performances. This comparative exercise reveals not only the complexity inherent in emotion detection within lyrical content but also the potential for distinct model advantages to manifest depending on the emotional framework being analyzed. The findings from this comparison provide pivotal insights, highlighting the intricacies of model selection and tuning in the field of emotion classification in song lyrics, and offering a data-driven foundation for future explorations aimed at enhancing the precision of multi-emotion detection models.

5. Discussion

5.1 Conclusions

This study presents a pioneering approach to multi-emotion classification in song lyrics, utilizing three distinct models: Naive Bayes, RoBERT-a, and Random Forest. Each model offers unique insights into the emotional spectrum of lyrics, underpinning the complex interplay of language and sentiment in music. The manual curation and labeling of the dataset stand as a testament to the depth of analysis and dedication to capturing the nuanced emotional landscape in songwriting. The findings demonstrate that each model holds specific strengths in deciphering and categorizing emotional content, highlighting the multifaceted nature of lyrical interpretation. This research not only advances our understanding of emotion classification in lyrics but also sets a precedent for future studies in the domain of computational musicology and affective computing.

5.2 Theoretical implications

The theoretical implications of this study are manifold. Firstly, it contributes to the existing body of knowledge in natural language processing (NLP) and emotion analysis by offering a novel application in the realm of song lyrics. The utilization of models like Naive Bayes, RoBERT-a, and Random Forest in this new context expands our understanding of their adaptability and effectiveness outside traditional text formats. Moreover, this research provides valuable insights into the emotional structure of lyrics, challenging and enriching prevailing theories in music psychology and linguistics about how emotions are conveyed and perceived in written form. It also opens avenues for exploring the synergy between linguistic cues and musical elements in evoking emotions.

5.2 Practical implications

From a practical standpoint, this research has significant implications for various industries. In the music industry, such an approach to emotion classification can aid in better categorizing songs for recommendation systems, enhancing user experience in music streaming services. It can also assist lyricists and composers in understanding emotional trends and audience preferences. Furthermore, in the field of mental health, the findings could be instrumental in developing therapeutic tools that use music and lyrics for emotional healing and expression. The methodology and models used can also inspire advancements in AI-driven content creation, where understanding emotional undertones is crucial for generating relatable and engaging material.

5.3 Limitations and future research

While this study marks a significant step in multi-emotion classification for song lyrics, it is not without its limitations. The primary constraint lies in the manual dataset curation and labeling, which, despite ensuring high-quality data, might contain inherent biases and limit the diversity of the dataset. Future research could address this by incorporating larger, more diverse datasets, possibly using crowd-sourcing for labeling to minimize biases. Additionally, exploring the integration of audio features with lyrical content could offer a more holistic approach to emotion classification in songs. Further studies could also experiment with more advanced NLP models or hybrid models that combine the strengths of the ones used in this study, to enhance accuracy and reliability in emotion classification.

References

Adam Dunbar, C. E. (2016). The threatening nature of “rap” music.

Darren Edmonds, J. S. (2021). Multi-Emotion Classification for Song Lyrics.

Downie, X. H. (2010). When lyrics outperform audio for music mood classification: A feature analysis.

Fried., C. B. (1999). Who’s afraid of rap: Differential reactions to music lyrics.

Gaojun Liu, Z. T. (2020). Research on multimodal music emotion classification based on audio and lyrics.

Genuis. (2009). Dohvaćeno iz https://genius.com/.

Hugging Face. (n.d.). Dohvaćeno iz Hugging face: https://huggingface.co/docs/transformers/index

Jia, X. (2022). Music Emotion Classification Method Based on Deep Learning and Improved Attention Mechanism.

Jiddy Abdillah, I. A. (2020). Emotion classification of song lyrics using bidirectional lstm method with glove word representation weighting.

Konstantinos Trohidis, G. T. (2011). Multi-label classification of music by emotion.

Lee, D. Y.-S. (2009). Music emotion identification from lyrics.

Loreto Parisi, S. F. (2019). Exploiting synchronized lyrics and vocal features for music emotion detection.

Lowe, S. (2022). *Hugging Face*. Dohvaćeno iz Hugging Face: https://huggingface.co/SamLowe/roberta-base-go_emotions

Lowe, S. (n.d.). *Hugging Face*. Dohvaćeno iz https://huggingface.co/datasets/go_emotions

Mayer, N. (2008). Rhyme and Style Features for Musical Genre Classification by Song Lyrics.

Mayer, R. (2011). Music Genre Classification by Ensembles of Audio and Lyrics Features.

Mihalcea, S. (2010). Annotating and Identifying Emotions in Text.

Rada Mihalcea, C. S. (2012). Lyrics,music and emotions.

Rémi Delbouys, R. H.-L. (2018). Music mood detection based on audio and lyrics with deep neural net.

Schubert, M. S. (2019). Negative emotion responses to heavy-metal and hip-hop music with positive lyrics.

Tsaptsinos, A. (2017). Lyrics-Based Music Genre Classification Using a Hierarchical Attention Network.

Xiao Hu, J. S. (n.d.). Lyric text mining in music mood classification.