Explicit Song Lyrics Analysis Using NLP Methods

KAI-YU LU, Northeastern University, USA
ARYAN KAKADIA, Northeastern University, USA
YIFEI SHANG, Northeastern University, USA

ACM Reference Format:
Kai-Yu Lu, Aryan Kakadia, and Yifei Shang. 2024. Explicit Song Lyrics Analysis Using NLP Methods. 1, 1 (December 2024), 14 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

In the current digital era, music permeates almost every aspect of daily life. However, certain genres and compositions can convey meanings that have adverse effects on listeners, particularly children or young adults, who may be exposed to lyrics containing aspects of violence, controlled sex, drugs, and alcohol. According to Primack et al. [12], adolescents consume an average of more than 14.7 hours of music per week, with a significant proportion of popular songs containing explicit references to sexual content, violence, and substance use. Consequently, such exposure could be related to earlier participation in risky behaviors, including substance abuse and premature sexual activity.

The American Academy of Pediatrics has highlighted the growing availability of music on digital platforms, and personal devices can often remove adolescents from parental supervision and exacerbate the impact of explicit lyrics [7]. In addition, lyrics that contain explicit content can reinforce harmful stereotypes, normalize dangerous behavior, and desensitize listeners to real-world consequences. In addition, music is often a vehicle for emotional expression and identity formation in adolescents, which can make them particularly susceptible to musical messages.

In this critical situation, it is essential to address the potential harm of explicit lyrics to youth development. Although existing parental advisory systems can offer some guidance, they are often criticized for their lack of specificity and inability to monitor content comprehensively. Therefore, our objective is to propose a novel language model to detect explicit lyrics based on the content of the lyrics to mitigate the influences of explicit lyrics on young adults. The model can not only detect basic keywords but also employ advanced natural language processing techniques to analyze implicit meaning and contextual nuances in lyrics. Through precise categorization of explicit content, the method can identify specific explicit themes, such as sexual content, violence, substance use, and language in song lyrics, thus facilitating interventions to reduce the unique risks posed by each subject and thereby negatively impact young audiences.

To better understand this study, the paper is organized as follows. We first discuss related work with regard to the classification of explicit lyrics, outlining past development achieved in natural language processing and multi-label classification. We then describe the data set in more detail, including the annotation process, corrections made to

Authors' Contact Information: Kai-Yu Lu, Northeastern University, Seattle, USA, lu.kaiy@noetheastern.edu; Aryan Kakadia, Northeastern University, Seattle, USA, kakadia.ar@northeastern.edu; Yifei Shang, Northeastern University, Seattle, USA, shang.yif@northeastern.edu.

Manuscript submitted to ACM

address ambiguities, and additional attributes used for correlation analysis. The section on experimentation provides the methodologies adopted in the study, including keyword-based initial annotation, bootstrapping for iterative refinement, and LLM-assisted verification through ChatGPT for the validation and improvement of the bootstrapping process. Lastly, the results are quantitatively and qualitatively analyzed with visualizations on performance metrics such as subset accuracy and per-category accuracy. The evaluation section contains not only a detailed comparison but also insight into the refinement of explicit categories and general effectiveness of the proposed approach.

2 Related Work

The paper "Detection of Explicit Lyrics in Hindi Music Using LSTM" investigated the application of Long Short-Term Memory (LSTM) networks for the automatic detection of explicit content in Hindi song lyrics [3]. The authors collected a dataset of 7,000 Hindi song lyrics, evenly categorized into explicit and non-explicit groups. For data preprocessing, the authors adopted methods of tokenizing the lyrics, removing punctuation and stop words, and converting the text into numerical representations by word embeddings. In the end, LSTM was trained to classify explicit lyrics. Although this model could have a satisfying 81.6% accuracy of detecting explicit lyrics, the performance of this study was still constrained by a small dataset, potential biases in the data labeling process, a narrow genre focus, and an inability to handle multilingual lyrics.

In the paper "Comparing Automated Methods to Detect Explicit Content in Song Lyrics," M. Fell et al. compared several automated methods for detecting explicit content in song lyrics [10]. They also addressed the challenge of detecting explicit content in song lyrics. By tagging explicit songs (songs that contain profanity, violence, or other inappropriate references or emotions), parents can identify content that may be unsuitable for children. From their research, M. Fell et al. (2019) concluded that while automated methods can assist in detecting explicit content, there is still a need for human oversight due to the subjective nature of the task.

The paper introduced a novel framework to improve the accuracy of sentiment classification of opinion targets [5]. The authors implemented bidirectional LSTM to generate weighted memory sentence representations and passed them into multiple attention layers to extract sentiment-related features. Afterward, these results would be combined with a GRU network to produce more accurate predictions. This approach achieved more outstanding results in understanding of complex sentence structures in four datasets than existing methods. However, the fixed number of attention layers and the complexity of the model would pose certain challenges in flexibility and scalability when facing varied contexts and datasets becoming larger.

T. Chen and C. Guestrin introduce the XGBoost model that is designed to handle sparse data with a novel tree learning algorithm and a weighted quantile sketch method [6]. It performs state-of-the-art classification and ranking tasks, including LambdaMART for ranking. It is also worth mentioning that XGBoost can handle sparse data, and its regularization techniques can avoid overfitting in significant measures. On the Higgs-1M dataset, XGBoost significantly outperformed both scikit-learn and R's GBM, and it can run more than 10 times faster than scikit-learn. This outstanding performance makes it suitable for lyric datasets, which often contain missing information due to varying song lengths, missing words, or inconsistent formatting.

M. Rospocher conducts experiments to explore the feasibility of using transformer-based language models (TLMS) such as BERT, DistilBERT, RoBERTa, XLNet, and DeBERTa to identify explicit content on over 800,000 lyrics datasets [14]. This study also demonstrates that self-attention mechanisms can perform well in capturing intricate linguistic patterns and long-range dependencies, which can further detect subtle explicit content more accurately than traditional

simple models. Additionally, with a limited amount of data, transformer-based language models can still maintain satisfying performance in detecting explicit content, demonstrating their adaptability in few-shot learning scenarios.

The paper "Adversarial and Domain-Aware BERT for Cross-Domain Sentiment Analysis" by Chunning Du et al. designed a post-training procedure, which contains the target domain masked language model task and a novel domain-distinguish pre-training task, fully utilizing language knowledge from the target domain and link the source and target in a self-supervised way [9]. For the problem of non-labeled data in the target domain, BERT would be pre-trained to distinguish whether the two sentences come from the same domain. Experiments on the Amazon reviews benchmark dataset, which employed a 5-fold cross-validation protocol, that is, in each fold, 1600 balanced samples were randomly selected from the labeled data for training and the rest 400 for validation, showed that this model got the average result in 90.12% in accuracy, remarkably outperforms state-of-the-art methods.

3 Dataset

3.1 Dataset Overview

The dataset used for initial model training is billboard-lyrics-spotify.csv, which is available on the Dataset Repository Link. It combines song lyrics and corresponding attributes like billboard chart data and audio features from Spotify. It is structured in a tabular format, with each row representing a unique song and columns capturing various attributes, including metadata, chart performance, lyrical content, and musical characteristics. The dataset's comprehensive nature facilitates research in natural language processing (NLP) aimed at detecting explicit lyrics and analyzing lyrical sentiment across popular music. The information within this dataset can be divided into four major areas:

- `Song Information`: The dataset contains basic information about the songs, like the song title, artist name, and release year. These fields can be used to categorize and filter based on specific time periods and artists.
- `Billboard Chart Data`: The dataset also contains information about the song's popularity. This includes information like the song's peak performance on the Billboard charts.
- `Lyrics Data`: The dataset also contains full song lyrics. This would enable natural language processing (NLP) and text analysis. Features such as word count and the presence of explicit content are also included in the dataset. This would allow us to examine trends in language use and the frequency of explicit content in popular music. By analyzing the lyrics, we can identify common themes, sentiments, or patterns that correlate with explicit songs.

This project aims to classify categories of song lyrics systematically; however, the existing Spotify dataset lacks labels for these categories. Consequently, the first step in our experiment was to develop a labeling category for the original Spotify dataset. The subsequent sections of this study will introduce the methodologies employed for categorizing the lyrics into specific classifications, which are sexual content, violence, substance use, and inappropriate language. This foundational work is essential for enhancing the dataset's utility and training this embryonic dataset for subsequent analyses.

3.2 Dictionary Preparation for Bootstrapping

Dictionaries are a significant part of natural language processing because they help machines understand and process human language understandably. They are made up of words or phrases used in daily speech or specific domains and play a key role in various NLP applications. Therefore, building a dictionary for the task with domain knowledge is indispensable.

To build a foundational dictionary, we begin with defining critical categories, including sexuality, violence, substance use, and language. Spontaneously, we referred to definitions from authoritative resources, notably the Entertainment Software Rating Board (ESRB). The ESRB is a self-regulatory organization tasked with assigning age and content ratings to video games across Canada, the United States, and Mexico. It aims to address concerns about video games that may contain excessively violent or sexual content. Additionally, the ESRB offers comprehensive definitions for explicit categories, and lyrics sometimes are involved in this rating system thus it is spontaneous to implement this rating system in this project.

3.3 Extend Dictionary by Word2Vec and Datamuse API

Due to the scarcity of words for categories and the fact that more vocabulary had a higher possibility of classifying the lyrics, it is crucial to further the dictionary. Initially, we employed Word2Vec, a neural network-based model that captures semantic relationships by mapping words into a continuous vector space [11]. The model effectively identifies words with similar contexts, which can help discover relevant terms for classification tasks. After filtering out words requiring contextual verification, the vocabulary size remained limited. To solve this problem, the Datamuse API is used, which provides rich word associations such as synonyms and modifiers, which allows us to generate additional vocabularies for each category.

3.4 Keyword Matching for Lyrics Analysis

To determine the categories of the lyrics accurately, each lyric was analyzed to match the keywords in each category. Moreover, to further accurately label the lyrics, a lemmatization strategy was adopted to normalize the morphology of the words. When a keyword indicating a particular category is identified in the lyrics, the relevant label is marked as true. In contrast, without relevant keywords, the corresponding labels remain unlabeled to reduce the possibility of false negatives. Therefore, this approach guarantees that only robust keyword-category associations are marked while minimizing redundant labels in cases where the associations are unclear. The overview of the labeled dataset is shown in Figure 1.

lyrics	Sexual	Violence	Substan	Languag	Sexual_wor	Violence_word	Substance_wor	Language_words
my daddy left home wher		T	T	T		gun, knife, kill	beer, booze	bitch
ow people moving out pe		T	T			gun	pill	
jeremiah wa a bullfrog wa		T				gun	wine	
saturday night i wa downt		T				gun	whiskey	
on a dark desert highway		T				knife, stab, kill	wine	
dearly beloved we are ga			T			kill	pill	
im all out of hope one mc	T				tease	murder	wine	
there a black man with a b			T			kill	pill	
your love is like bad medi			T			attack	drug, pill	
in the time of chimpanzee			T			kill	cocaine	
im an indian outlaw half cl			T			kill	drug, wine	
at these up late time hard	T	T	T	T	ass	gun	pill	motherfucker, fuck, shit, bitch, fuckin, ass
uh junior m a f i a uh yeah	T	T		T	ass	gun	wine	nigga, fuck, shit, fuckin, nigga, ass
when i step up in the plac	T	T	T	T	ass	gun	weed	nigga, motherfucker, fuck, shit, nigga, ass
do you wanna ride in the l	T		T	T	dick, sex	slay	weed	nigga, motherfucker, shit, bitch, nigga
one two three four oh you		T				gun	beer	
i know a dude named jim	T		T	T	ass, dick, sex	kill	weed	nigga, motherfucker, fuck, shit, nigga, ass, fucker
artist heavy d album wate	T		T		sex, sexy	kill	weed, wine	
ghetto supastar that is wh		T	T			gun	drug	
cadillac grill cadillac mill cl	T		T	T	ass, dick	kill	drug, liquor	nigga, nigga, ass
wheres my snare i have nc	T	T	T	T	ass, panty	gun	pill	shit, bitch, fuckin, ass
move bitch get out the wa	T		T	T	ass, dick, sex	punch	pill	nigga, motherfucker, fuck, shit, bitch, fuckin, nigga, ass
roll out i got my twin gloc	T	T	T	T	ass, dick	knife	weed	nigga, fuck, shit, bitch, nigga, ass
go go go go go go go sh	T	T	T	T	ass, sex	knife	weed, drug	nigga, motherfucker, fuck, bitch, fuckin, nigga, ass

Fig. 1. Result of the labeled dataset after matching keywords for lyrics segment

4 Experimentation

4.1 Bootstrapping for Iterative Label Refinement

Before training the model, a bootstrapping strategy with a sliding window mechanism for segment extraction was implemented to enhance the classification of lyrics into explicit content categories. For each keyword detected in the lyrics, extract a fragment of 25 words and position the keyword in the center of the window. This design ensures that the semantic context surrounding the keyword is preserved and captures both direct and indirect associations with the explicit topic.

Each segment was compared to pre-computed embeddings of category-specific keywords using Sentence-BERT [13] (SBERT), and a semantic similarity score was calculated. If any segment exceeded the current threshold (initially set at 0.62 and reduced incrementally by 0.02 per iteration to a minimum of 0.56), the lyric was labeled True for the corresponding category. These high-confidence segments were added to the training set as new positive examples.

Each segment of the lyrics is compared with a pre-calculated keyword embed for a specific category using SBERT to calculate a semantic similarity score. When any part exceeds the established threshold (initially set at 0.62, then reduced by 0.02 with each subsequent iteration to a minimum of 0.56), the corresponding lyric is marked as True. These segments are determined to be of high confidence and then added to the existing training set as new positive examples.

To ensure that the model remains adaptable and avoids overfitting, SBERT fine-tunes itself by using an updated training set after each round, which enables the model to gradually increase its understanding of the explicit content associated with each category. In addition, the use of sliding window mechanisms and dynamic threshold adjustments ensures efficient identification of local keyword matches while also capturing broader contextual meaning.

The preprocessed lyrics in the original dataset lack a clear structure, which made it hard to classify categories like Substance and Violence accurately because these categories might rely on context, where meaning is shaped by surrounding text rather than just only keywords. To improve this, and because of the complex context surrounding Substance and Violence and numerous lyrics to be further verified, we introduce Large Language Models and manual verification in this stage. This method is more efficient and reduces labor costs than aimless labeling.

4.2 LLM-Assisted (ChatGPT) Verification for Bootstrapping in Lyrics Classification

To further assess the reliability of the methods based on bootstrapping for classifying song lyrics, ChatGPT 4 was used as an auxiliary validation tool at this stage [1]. Carefully designed task-specific prompts incorporating CoT reasoning were used to generate categorical outputs from the model based on the same dataset processed through bootstrapping. The CoT prompts allowed the model to generate intermediate reasoning steps, which provided a clearer and more interpretable comparison of its outputs with those from the generated bootstrapping method.

Results from both methods were compared, and discrepancies among samples were manually verified to ensure accuracy. This dual-method approach not only made for an external reference with regard to the results found by bootstrapping, but it also showed which points in the bootstrapping needed further semantic capabilities and contextual reasoning. Blending the advanced contextual reasoning of the language model, enabled through CoT reasoning, with human verification underlines the importance of large machine learning tools combined with human oversight to achieve robust yet precise classification systems. The overall flowchart for this LLM-assisted verification can be referred in Figure 2.

After the dictionaries were expanded and labeling was done through the dictionaries and the language model, the final refinement of the dataset was done, totaling 1,234 lyrics. Of those, 1,034 lyrics had at least one explicit category,

while the remaining 200 lyrics were classified as non-explicit. This collection forms the distilled dataset, which is the training set described in Section 4.3.3. The term "distilled dataset" is used to indicate that it is refined since it best represents explicit content for each category.

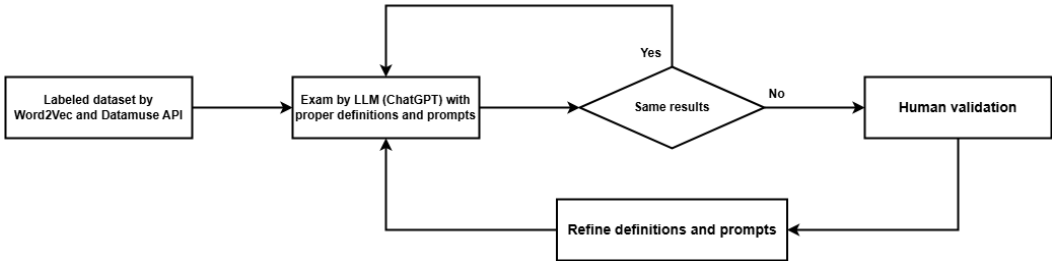

Fig. 2. Overall flowchart for the LLM-assisted verification

4.3 Training Models

4.3.1 Overview of Selected Models. In this experiment, three transformer-based models, BERT, BigBird, and Longformer, are explored for multi-label lyrics classification. The reason for selecting the above-mentioned models relies on the different approaches followed by these models in handling textual input of variable length. To enable the handling of longer sequences, the BERT model is extended into a Dual-BERT setup by analyzing the starting and ending parts of the lyrics separately. BigBird's sparse attention mechanism, in conjunction with Longformer's sliding window attention, enabled the efficient processing of long sequences without sacrificing contextual richness. The two models were optimized to particularly address the task of categorizing lyrics into four specific categories: Sexual, Violence, Substance, and Language.

4.3.2 K-Fold Cross-Validation. To ensure strong evaluation and counter-overfitting, a 5-fold cross-validation strategy was performed for each of the three models. The dataset was divided into five stratified folds to ensure balanced label distributions were maintained across the folds. At each iteration, four of the folds are used in training, and the rest are used for validation. This ensures that every single data sample contributes to training and validation, hence providing a more holistic assessment of model performance.

Performance metrics were recorded for each fold, including accuracy, precision, recall, and F1 score, and then averaged across all folds. It not only helped in better understanding the strengths of each model but also allowed comparison on fair terms across different architectures.

4.3.3 Model Training and Implementation. The three models, BERT, BigBird, and Longformer, are carefully fine-tuned to accommodate the unique challenges of multi-label lyrics classification. Each of these models, by the underlying mechanism of attention and architectural design, exhibited some benefits.

BERT was modified to a Dual-BERT setup to account for its input token length limit, which is capped at 512. Lyrics longer than this were divided into two halves: the start and end of the lyrics.

The above segments were separately passed through a BERT model with the same shared parameters. The pooled outputs of the two segments were then concatenated and passed to a classification head. This architectural design allowed Dual-BERT to retain information from both the starting and ending parts of the lyrics and thus capture critical contextual signals that are often spread over a song.

BigBird used a sparse attention mechanism that combined local, global, and random attention to deal with long sequences efficiently. This helped BigBird maintain both local dependencies and a global context—something especially suited for lyrics, where key phrases might be far apart from each other. BigBird supported up to 1024 tokens in the implemented experimental configuration to ensure that there was consistency in the length of input across models.

Longformer used a sliding window attention mechanism to capture local dependencies, but it selectively attended to global tokens in order to encapsulate the overall context. Such integration allowed Longformer to process lyrics as a cohesive sequence while simultaneously retaining computational efficiency. This turned out to be especially effective at picking up the repeating phrases or keywords traversing different parts of the lyrics, thus fine-tuning its ability to correctly classify subtle content. These models are trained using Hugging Face's Trainer framework, which provides consistent training pipelines. Differences in attention mechanisms and input handling have further underscored their adaptability to the challenges of lyrics classification.

4.4 Evaluation Metrics

The performance of each model was evaluated using overall and category-specific metrics, conveying all strengths and limitations of the models. The main metrics included:

- **Subset Accuracy:** This was a strict measure, as it assessed the ratio of instances where all labels predicted exactly matched the ground truth, hence being a strong indicator of thorough performance in multi-label classification. The formula is defined as:

$$\text{Subset Accuracy} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(\hat{Y}_i = Y_i)$$

  Where:
  - N: Total number of samples.
  - $\hat{Y}_i$: Predicted label set for the i-th sample.
  - Y_i: True label set for the i-th sample.
  - $\mathbb{I}(\cdot)$: Indicator function, which equals 1 if the condition is true, and 0 otherwise.

  Subset accuracy is highly sensitive to even small classification errors, making it an excellent metric for assessing the robustness of multi-label models.
- **Per-Category Accuracy:** The individual accuracy for each category, namely, Sexual, Violence, Substance, and Language, was computed separately, thus showing the respective strengths and limitations of the models concerning specific content types. The formula is:

$$\text{Accuracy}_{\text{category}} = \frac{\text{TP}_{\text{category}} + \text{TN}_{\text{category}}}{\text{TP}_{\text{category}} + \text{TN}_{\text{category}} + \text{FP}_{\text{category}} + \text{FN}_{\text{category}}}$$

  Where:
  - $\text{TP}_{\text{category}}$: True positives for the category.
  - $\text{TN}_{\text{category}}$: True negatives for the category.
  - $\text{FP}_{\text{category}}$: False positives for the category.
  - $\text{FN}_{\text{category}}$: False negatives for the category.

  Per-category accuracy highlights the model's capability to classify specific categories accurately, offering insights into areas requiring improvement.

- **Precision, Recall, and F1 Score:** These performance metrics have been used for the evaluation of the true positives' rates against false positives rates. They allowed a much deeper look into how the models performed concerning the imbalanced datasets for each category.
- **Confusion Matrix:** The results of each model were visualized to identify patterns of misclassification across categories. This helped us to understand where the models were struggling and where their predictions were in good agreement with the ground truth. By integrating these metrics, the evaluation framework has managed to contain the complex performance of each model, exemplifying their ability to learn the semantic and contextual features inherent in lyrics.

5 Results

5.1 Results Overview and Comparative Analysis

This section mainly involves the results of several models used in this project, and these models include BERT [8], BigBird [2], and Longformer [4] with four classification categories: Sexual, Violence, Substance, and Language. Additionally, it is noted that the baseline performance is calculated from majority class predictions. The overall results for each model are summarized in Table 1, while a detailed confusion matrix is presented in Figure 3, Figure 4, and Figure 5. The evaluation methods are mentioned in the previous Section 4.4.

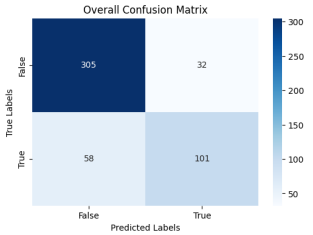

Fig. 3. Overall Confusion Matrix for BERT

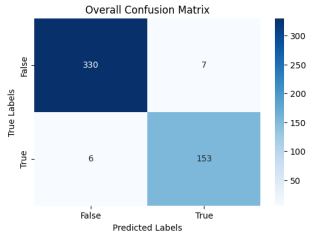

Fig. 4. Overall Confusion Matrix for BigBird

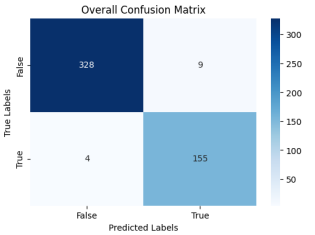

Fig. 5. Overall Confusion Matrix for Longformer

Model	Sexual	Violence	Substance	Language	Subset Accuracy
Baseline (Majority)	52.51%	83.36%	80.36%	65.28%	-
BERT	70.97%	87.90%	84.68%	83.87%	41.13%
BigBird	95.16%	97.58%	97.58%	99.19%	91.13%
****Longformer****	**95.16%**	**97.58%**	**98.39%**	**98.39%**	**91.94%**

Table 1. Performance comparison of different models

5.2 Inference on the Original Dataset Based on Explicit Categories

This dataset originally consisted of 5,566 lyrics. Once the entries with empty lyrics had been removed, a total of 5,492 usable items remained in the dataset. Each lyric was run through inference by the use of the best Longformer model, explained in Section 6.1, and each lyric was labeled as explicit or non-explicit. Among the inferred lyrics, the number of items containing at least one explicit category is 1,925 (35.05%), while the remaining 3,567 items (64.95%) are non-explicit.

These explicit categories under analysis indeed show the diversity of content in these datasets. In both the original and explicit datasets, there are 973 associated with sexual content, 318 with violent themes, 401 referencing substance use, and 915 with explicit language. The explicit dataset contains 1,925 lyrics and emphasizes explicit content (at least one explicit category). In this subset, 50.55% of the content is sexual, 16.52% of it is violent, 20.83% refers to substance use, and 47.53% contains explicit language.

The dataset with 5,492 lyrics includes all categories with the majority of non-explicit entries. Sexual content appeared in 17.72% of the lyrics, violent theme expressions in 5.79%, substance use in 7.30%, and explicit language in 16.66%. While the explicit dataset narrows its focus to explicit themes, the dataset now depicts an overview of lyrical content in all aspects.

Figures 6 and 7 present the distribution of these explicit categories. Figure 6 presents the proportions in the explicit dataset, while Figure 7 presents the broader distributions found in the original dataset. The contrast between these figures is striking, illustrating how the explicit dataset focuses on the explicit content of the original dataset.

Within the dataset comprising 5,492 usable items, a total of 5,330 items were designated as either explicit or non-explicit, leaving 162 items unclassified. The original dataset demonstrated considerable deficiencies in its labeling accuracy, especially in the underrepresentation of explicit lyrics. Among the 5,330 labeled lyrics, merely 515 were classified as explicit, in contrast to 4,815 that were categorized as non-explicit. After the re-evaluation of this project, it was found that 1,925 lyrics were classified as explicit and 3,567 as non-explicit. Moreover, 1370 explicit lyrics and 26 lyrics are mismatched in the original dataset (originally unlabeled lyrics are excluded). More in-depth descriptions of the discrepancies in the case studies will be given later.

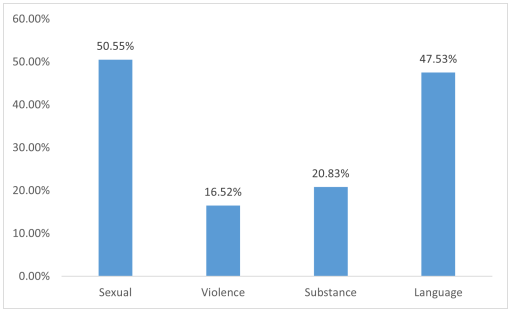

Fig. 6. Distribution of Explicit Categories in the Explicit Dataset

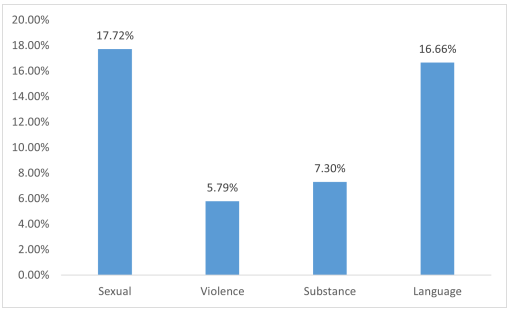

Fig. 7. Distribution of Explicit Categories in the Original Dataset

6 Evaluation

6.1 Model Evaluations on Distilled Dataset

We evaluated the effectiveness of these models in the task of multi-label classification of song lyrics in the distilled dataset mentioned in Seciton 4.2, focusing on explicit content categories such as Sexual, Violence, Substance, and Language. The confusion matrices of the Dual-BERT, BigBird, and Longformer models showed quite different patterns for performance, outlining the specific strengths and limitations of each model in this respect.

The Dual-BERT architecture uses two BERT encoders to handle up to 1024 tokens by dividing the input into two parts and concatenating their outputs. This design tackles the single-BERT limitation of only being able to handle 512 tokens, allowing Dual-BERT to capture a much wider context in longer lyrics. As shown in its confusion matrix, this extension

reduces false negatives, especially for subtle explicit content. Dual-BERT performs moderately well with per-category accuracy scores of 70.97% for Sexual content and 87.90% for Violence, along with a subset accuracy of 41.13%. These results, while better than more basic baselines, underline the intrinsic challenges of multi-label classification in handling diverse and extensive lyrical content.

In contrast, BigBird does well in all categories by using its sparse attention mechanism; thus, it efficiently manages sequences of up to 4096 tokens. The confusion matrix of BigBird reduces the misclassifications significantly with respect to Dual-BERT. Its accuracy per category for both Violence and Substance exceeds 97%. Its subset accuracy of 91.13% is very strong in maintaining consistency on making predictions with the ground truth. Results show that BigBird outperforms in capturing long-range dependencies and contextual associations, rendering it effective for the tasks at hand.

Longformer outperforms both Dual-BERT and BigBird, where it leverages a sliding window attention mechanism that combines computational efficiency with the ability to model local and global contextual relationships within lyrics. Its confusion matrix shows minimal errors with accuracy above 98% for Substance and Language categories. Longformer's subset accuracy of 91.94% is only slightly better than BigBird, sealing its place as the strongest model for this task. Per-category accuracy further accentuates BigBird and Longformer over Dual-BERT. Both higher-order models excel in the Sexual category, each in their respective measure—each at 95.16% accurate. Longformer bested BigBird in the Substance category with 98.39% versus 97.58%. These findings highlight the critical role of refined attention mechanisms in detecting subtle and implicit explicit content.

Baseline performance, contrasted with majority-class predictions, puts into perspective the limitations of such strategies, particularly in challenging classes such as Sexual (52.51%) and Language (65.28%). Subset accuracy, known to be a strict metric for multi-label classification, better illustrates the superiority of BigBird and Longformer in making their predictions align with the ground truth. While Dual-BERT has an advantage due to its higher contextual capacity, it achieves lower subset accuracy because of its split input architecture, which may cause a loss of coherence in larger lyrical contexts. In summary, the results underline the enormous benefits of advanced transformer-based architectures for explicit lyrics classification. While Dual-BERT represents a pragmatic improvement over vanilla BERT for longer input sequences, BigBird and Longformer outclass it by handling long-range dependencies and contextual subtleties effectively. Of this cohort of models, Longformer has proven to be the most robust, with special prowess in semantic understanding and correct multi-label classification. These results show that architectures optimized for deep contextual knowledge are necessary for dealing with the unique difficulties presented by explicit lyrics classification.

6.2 Data Analysis on Original Dataset

The initial dataset, with 5,492 song lyrics after the preprocessing process, has a more diverse and comprehensive collection of lyrical content compared to the explicit dataset. Among this collection, 35.05% (1,925 lyrics) are identified to include at least one explicit category, while 64.95% (3,567 lyrics) are labeled as non-explicit. Such labeling shows the larger and more general characteristics of the original dataset, covering more diverse lyrical content.

A look at the explicit categories in the original dataset suggests that sexual content is the most cited explicit theme, making up 17.72% of the lyrics. Thus, it has a big focus in the explicit subset. The second would be explicit language, coming in at 16.66% of the lyrics; this reflects the wide use of vulgar or suggestive words to create a style or for audience appeal. Citations about substance use make up 7.30%, reflecting the widespread use of drugs, alcohol, or tobacco themes in music. In comparison, while violent themes are the least common of explicit categories, they still feature in 5.79% of lyrics, indicating that their presence in musical expression, though not dominant, is still very much there.

Among the clearly categorized lyrics—1,925 items from the original data set—sexual content and explicit language are the most striking features, together making up a large portion of the explicitly classified lyrics. Their relatively high frequency indicates their importance to explicit lyrical expression. On the other hand, substance use and themes of violence are less frequent but still hold a substantial place in the spectrum of explicit themes.

The results show that sexual content and explicit language are the most frequent explicit themes, dominating both the explicitly classified subset and the overall explicit content in the data set. Other themes, such as substance use and violence, appear less in frequency but contribute to a wider range of explicit themes found in song lyrics.

Figures 7 and 6 display a graphical view of how explicit categories are distributed, clearly showing their respective proportions. The analysis puts into light the very high rate of sexual content and explicit language in the original dataset, as well as their large representation in the explicit subset.

Figure 8 reveals important trends in the explicit lyrical themes in conjunction with their musical environment. The results of the correlation analysis show that linguistic content is most strongly related to popularity (r=0.26) and loudness (r=0.25), suggesting that the use of explicit language may act to enhance commercial appeal and intense expression. Moreover, the positive relation of sexual content to danceability (r=0.23) indicates its frequent use in rhythmic, dance-oriented genres. The inverse relationship observed between instrumentalness and all explicit categories, especially Language (r=-0.18), indicates that explicit content primarily manifests in compositions that emphasize vocals.

Figure 9 presents the results from a joint probability analysis, showing Sexual and Language to be the most frequent categories (0.505 and 0.475, respectively) that often co-occur (0.228), whereas Violence appears to be less common and hardly combines with other explicit themes. This would suggest a trend where explicit content in modern music is more likely to concern sexual themes and language than violence or substance use.

The relatively modest correlation coefficients shown in Figure 8 suggest that explicit content and musical attributes retain much independence, which may indicate that explicit themes might serve divergent artistic purposes rather than being anchored to specific musical characteristics. These findings add to our understanding of how explicit content functions within modern musical composition and might inform approaches to content analysis in music.

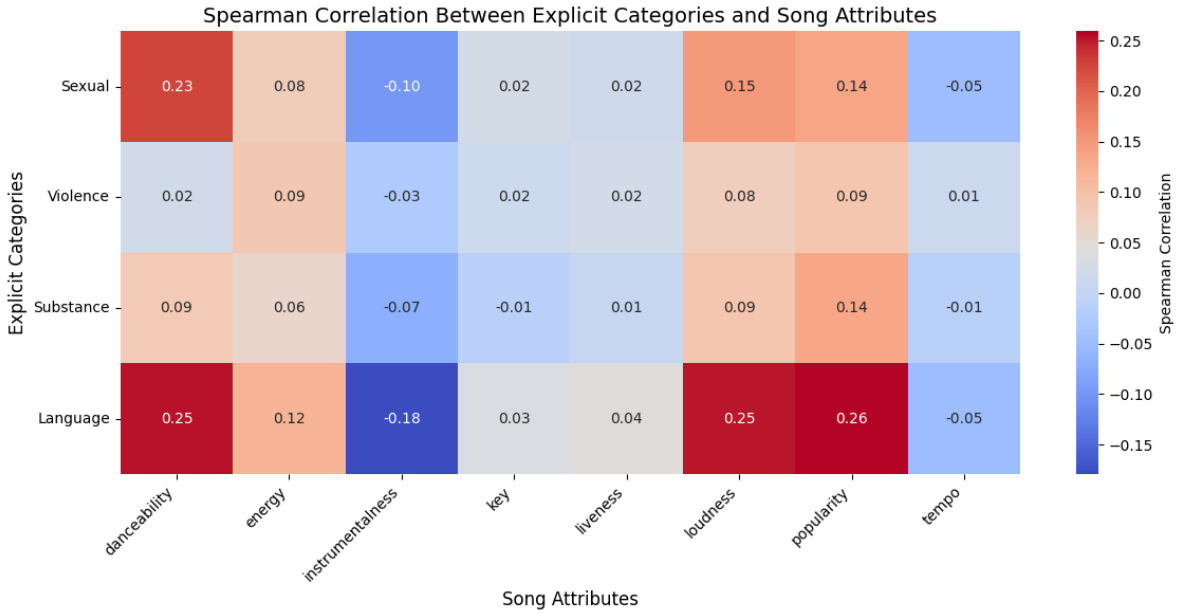

Fig. 8. Correlation Between Explicit Categories and Song Attributes

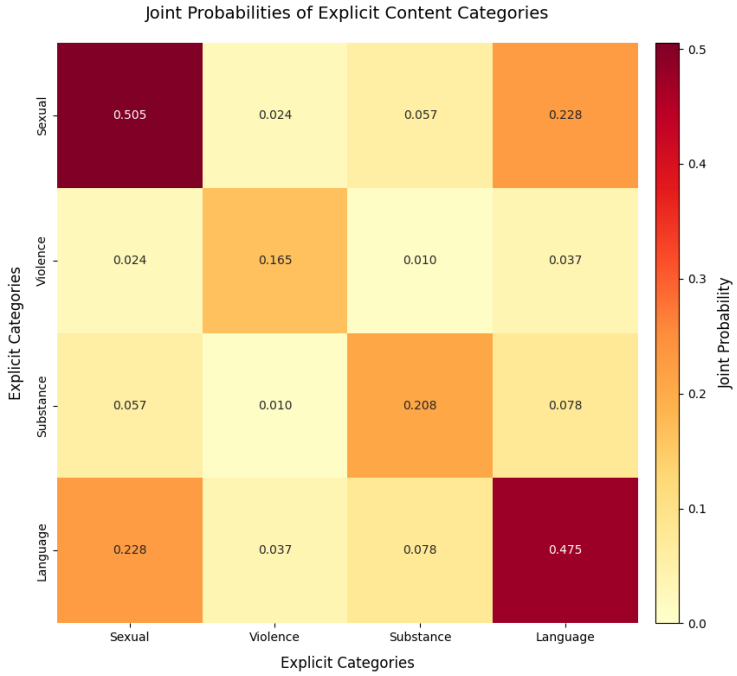

Fig. 9. Joint Probability of Explicit Categories

7 Conclusin

This article presents a comprehensive study on the classification of explicit content in song lyrics using NLP methods. The authors underline the negative influence that explicit lyrics can have on young listeners and propose a new language model to classify lyrics into four classes: sexual content, violence, substance use, and inappropriate language. The dataset used for this analysis is billboard-lyrics-spotify.csv, which combines song lyrics with associated features such as chart performance and audio features.

They implement various techniques to make an end-to-end classification model, starting with preparing dictionaries, keyword matching, bootstrapping to expand labels iteratively, and large language model-assisted validation through ChatGPT. Three transformer-based models (BERT, BigBird, and Longformer) are fine-tuned and evaluated in a 5-fold cross-validation setting. The results of experiments show strong performance from the Longformer model, capable of catching long-range dependencies and subtle contextual details; it also achieves the highest subset accuracy of 91.94% and passes over 95% in per-category accuracy metrics.

Further analysis of the original data set suggests that sexual content and explicit language are the most frequently occurring explicit categories, while violence and substance use are less well-represented. The study further explores the relationships between explicit categories and song features, finding strong associations between explicit language with popularity and loudness and positive associations between sexual content and danceability.

This study will provide a means of detecting explicit content in song lyrics using state-of-the-art techniques in natural language processing. The proposed model, developed based on the Longformer architecture, shows high performance in classifying explicit genres of lyrics. Results have high importance, presenting the harmful influence that explicit lyrics have on teenagers and basically showing an understanding of the spread and characteristics of explicit musical content. Future studies will, therefore, concentrate on incorporating acoustic features that can be used to broaden the analysis to multilingual song lyrics for better generalizability and practicality of the model.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[2] Shima Asaadi, Saif Mohammad, and Svetlana Kiritchenko. 2019. Big BiRD: A Large, Fine-Grained, Bigram Relatedness Dataset for Examining Semantic Composition. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, Minneapolis, Minnesota, 505–516. https://doi.org/10.18653/v1/N19-1050

[3] Nomi Baruah, Amlan Jyoti Kalita, Anugya Gogoi, Madhuzya Bezbaruah, Nikesh Prasad, Vishma Pratim Das, and Rituraj Phukan. 2023. Detection of Explicit Lyrics in Hindi Music Using LSTM. In *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*. 1–5. https://doi.org/10.1109/ICCCNT56998.2023.10308139

[4] Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The Long-Document Transformer. *arXiv:2004.05150* (2020).

[5] Peng Chen, Zhongqian Sun, Lidong Bing, and Wei Yang. 2017. Recurrent Attention Network on Memory for Aspect Sentiment Analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Martha Palmer, Rebecca Hwa, and Sebastian Riedel (Eds.). Association for Computational Linguistics, Copenhagen, Denmark, 452–461. https://doi.org/10.18653/v1/D17-1047

[6] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, California, USA) *(KDD '16)*. Association for Computing Machinery, New York, NY, USA, 785–794. https://doi.org/10.1145/2939672.2939785

[7] Council on Communications and Media. 2009. From the American Academy of Pediatrics: Policy statement–Impact of music, music lyrics, and music videos on children and youth. *Pediatrics* 124, 5 (Oct. 2009), 1488–1494.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[9] Chunning Du, Haifeng Sun, Jingyu Wang, Qi Qi, and Jianxin Liao. 2020. Adversarial and Domain-Aware BERT for Cross-Domain Sentiment Analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 4019–4028. https://doi.org/10.18653/v1/2020.acl-main.370

[10] Michael Fell, Elena Cabrio, Michele Corazza, and Fabien Gandon. 2019. Comparing Automated Methods to Detect Explicit Content in Song Lyrics. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, Ruslan Mitkov and Galia Angelova (Eds.). INCOMA Ltd., Varna, Bulgaria, 338–344. https://doi.org/10.26615/978-954-452-056-4_039

[11] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781 [cs.CL] https://arxiv.org/abs/1301.3781

[12] Brian A Primack, Erika L Douglas, Michael J Fine, and Madeline A Dalton. 2009. Exposure to sexual lyrics and sexual experience among urban adolescents. *Am J Prev Med* 36, 4 (April 2009), 317–323.

[13] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, Hong Kong, China, 3982–3992. https://doi.org/10.18653/v1/D19-1410

[14] Marco Rospocher. 2022. On exploiting transformers for detecting explicit song lyrics. *Entertainment Computing* 43 (2022), 100508. https://doi.org/10.1016/j.entcom.2022.100508