Recommending Appropriate Lyrics to Youngsters

Understanding the Presence of Inappropriate Content in Music Lyrics: Insights for Children's Recommender Systems

Jasper Heijne[1]

Supervisors: Sole Pera[1]**, Robin Ungruh**[1]

[1]**EEMCS, Delft University of Technology, The Netherlands**

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 22, 2024

Name of the student: Jasper Heijne
Final project course: CSE3000 Research Project
Thesis committee: Sole Pera, Robin Ungruh, Julian Urbano Merino

Abstract

Recommender systems play a considerable role in the consumption of music, also for children. Children are easily influenced, inappropriate song lyrics can negatively impact children's behaviour and personality, by teaching inappropriate language or harmful biases. We argue for a need for recommenders to protect children from inappropriate content. Recommender systems should consider what content is and is not well-suited for children. To guide future design and development of recommender systems for children, this research probes the music lyrics of Pop, Rock, Country, Rap, and R&B for inappropriate content. We achieve this by an empirical analysis using four existing algorithms to scan different facets of inappropriate content in the lyrics in a dataset of $37,993$ songs from the Genius database. The outcome of this empirical exploration reveals that Rap and R&B are the most inappropriate music genres and advises to be cautious when recommending these genres to children. Results also show that inappropriate content is highly prevalent in music. We address the need for a filter for recommender systems, to filter inappropriate songs for children.

1 Introduction

Children are vulnerable to the influences of music [6, 17]. These influences can negatively impact children's behaviour and personality, e.g. when children copy inappropriate behaviour mentioned in lyrics, repeat inappropriate words and sentences, or experience negative emotions like depression. These influences are even more impactful since children are in the developmental phase of their lives [5, 20]. Nowadays, there is a large selection of songs available online to children that can impact them negatively. Recommender systems are widely used to present new songs from this large selection of possibilities. These selected songs are based on training data, from this data, the algorithm learns what songs are enjoyed by certain users. With this knowledge, the system can recommend songs to other users. If this training data contains inappropriate content, the model might reflect this tendency and suggest inappropriate content. The lack of supervision over recommendations allows recommender systems to suggest harmful items to their users. For example, violence, stereotypes, or hate speech can make their way to children and can have bad consequences on their behaviour [15, 19, 23]. This is how recommender systems could negatively influence children in the developing phase of their lives. Therefore, there is a need for these systems to avoid suggesting inappropriate content to children. Currently, binary labels exist indicating inappropriate content in songs. However, the degree to which they capture different types of harm and how effective they are is unsure. An overview of inappropriate music content is necessary to inform recommender systems well.

This paper creates an in-depth overview of inappropriate content in the music genres: Pop, Rock, Rap, R&B, and Country. We discuss the prevalence of inappropriate content based on four facets: curse words, profanity, offensive speech and hate speech. We quantify the presence of each of those categories in existing lyrics using a different model per facet to scan a total of 37993 songs. This results in a clear overview of inappropriate content in the five music genres. These facets create a profile per genre of inappropriate content in the lyrics. Based on previous work, these profiles are related to music recommender systems for children to show how the content could influence them. This answers the main research question of this paper:

What is the prevalence of inappropriate content in lyrics of modern Pop, Rock, Country, Rap, and R&B music and how does this influence music recommender systems for children?

This research will contribute to the research field of music recommender systems for children, by advising future recommender systems on inappropriate content in lyrics in Pop, Rock, Rap, R&B and Country music. This information will help these recommender systems to consider if songs will negatively influence children with inappropriate content.

2 Related Work

Bandura's social learning theory suggests that children learn new behaviours by observing and imitating others [7]. This theory complies with research on the influence of music on children, which concludes that music can negatively influence children. Youth listening to metal music, for example, are more likely to show reckless behaviour, in the form of reckless driving, shoplifting, vandalism, reckless sexual behaviour, and drug use [6]. Work by Martin et al. [17] shows similar results claiming: "Significant associations appear to exist between a preference for rock/metal and suicidal thoughts, acts of deliberate self-harm, "depression", "delinquency", drug taking, and family dysfunction." This shows music can badly influence children.

It is shown that recommender systems do not filter inappropriate content well. Children are likely to encounter inappropriate content on YouTube due to its recommender system [18]. YouTube also amplifies extremist content in their recommendations [23], this shows that recommender systems can recommend inappropriate content. We see from previous work that inappropriate songs can influence children negatively, so music recommender systems should consider what content to recommend to children.

To identify possibly harmful inappropriate content in songs, the music industry adopted a label to show a song contains explicit content. Since 1987 albums have had a label with "parental advisory" [16] to show that the content may be explicit. Spotify also uses an "Explicit"[1] label for songs. This label allows the user to avoid all music labeled explicit but the functionality of the application is not nuanced in any way. A song labelled explicit is always avoided if the user chooses to exclude explicit music, without considering in what way the song is explicit. This is partially because

[1] Explicit label Spotify: https://support.spotify.com/us/article/explicit-content/

explicit content in songs is not structurally represented. The current "parental advisory" and "Explicit" labels are binary, giving an unnuanced representation of the inappropriateness of the song. Recommender systems can only extract a limited amount of useful insights from this binary data.

The detection of inappropriate content in music was the subject of previous research. Research by Frisby and Behm-Morawitz [12] has shown that many popular music of recent years contains explicit lyrics. However, this research was done manually and therefore on a small sample of 100 songs, not capturing the wide range of music and genres that children listen to. Other studies have been performed on analysing lyrics of music to check for inappropriate content, e.g. in the form of misogyny [4], violence [13], or drugs and alcohol [14]. These studies all found that inappropriate content is highly prevalent in popular music. The research on drugs, alcohol, and sex [14], found that those forms of inappropriate content were most prevalent in Rap and R&B music. All mentioned work is however manually executed on a relatively small sample, missing the complete overview of the broad musical spectrum children listen to. The mentioned inappropriate content detection studies also do not relate their findings to recommender systems.

Research centred on the computer science side of inappropriate content in lyrics often researches methods to scan for inappropriate content [9, 10]. These papers conclude with a statement explaining that the proposed method works. However, research studying the best algorithm for scanning inappropriate content refrains from scanning music lyrics and does not inform recommender systems. Thus, the conclusions cannot be used directly by recommender systems. Other work [21] focuses on the differences in performance of multiple machine learning algorithms. The results show what model works best for scanning toxicity, which could help recommender systems adopt the right model to filter inappropriate content. However, this work gives no further information on the musical landscape and does not mention recommender systems explicitly.

The main gap remaining in current research is a large-scale overview of inappropriate content in music lyrics. Currently, any information on inappropriate content is not related to children's recommender systems. The work presented in this manuscript fills these gaps in the research field and informs future recommender systems for children on inappropriate content.

3 Experimental Setup

In this Section, we describe the experimental setup, along with empirical explorations conducted to analyze the prevalence of inappropriate content in lyrics. The scripts for data processing and analysis can be found in this paper's repository [2]. In sections 3.1 and 3.2 we explain what data is used and why it is used. In section 3.3 we explain what models are used to scan inappropriate content in the lyrics and why they are used. In section 3.4 we explain how the models are applied to get results.

3.1 Dataset

The database we use for song lyrics and their respective genres is the Genius lyrics database [3]. This database is well suited for this experiment since the data originates from the Genius website [4], which is kept up-to-date by the large user base. The database was extracted from this website from September 2019 to January 2020, meaning the data is somewhat recent. The user base of Genius also has a substantial proportion of young users [3], which means that the songs on the website are likely interesting to a young audience.

From the Genius database, we use the lyrics for the 37993 songs; along with the genre tags provided for each respective song. This allows sorting all lyrics by their respective genre to get results on inappropriateness per genre.

3.2 Genres

The Genius database uses tags to represent the genres of songs. There are 5 main tags: Pop, Rock, Country, Rap, and R&B. Every song in the database has at least one of these five main tags, meaning all songs in the database can be scanned for inappropriate content using these tags as genres.

The number of songs can be found in Table 1. The proportion of Rap songs is highest as Genius has Rap music as its main focus. Pop, Rock, and R&B all have a similar number of songs. The Country genre has substantially fewer entries, but they are still included to cover the entire lyrics database.

Genres	**Number of Songs**
Rap	30,050
Pop	5,223
R&B	4,537
Rock	3,320
Country	164
Total	**43,294**

Table 1: Number of songs per genre in the Genius dataset used in our exploration.

Each song in the database can be categorized under one or multiple genres from the five main genres. Table 2 shows how many songs overlap for every pair of genres. The numbers on the diagonal are the total number of songs per genre.

	Pop	**Rock**	**Country**	**Rap**	**R&B**
Pop	5223	763	100	1201	1263
Rock	763	3320	42	291	90
Country	100	42	164	28	2
Rap	1201	291	28	30050	2442
R&B	1263	90	2	2442	4537

Table 2: Overlap counts of songs between genres. Each entry describes how many songs are in both corresponding genres.

[2] This paper's repository: https://github.com/JasperHeijne/research_project

[3] Genius database: https://www.cs.cornell.edu/~arb/data/genius-expertise/

[4] Genius website: https://genius.com/

3.3 Models for Inappropriate Content Detection

To create an overview of inappropriate content in music lyrics four models are used. Each of the models scans for a different type of inappropriate content: curse words, profanity, offensive speech, or hate speech. These four kinds of inappropriate content form a profile per genre and allow comparing genres on their prevalence of inappropriate content.

All models scan a different aspect of inappropriateness, leading to a more detailed overview of parts of inappropriate content. Such specific information allows for a more detailed comparison between genres and a broader overview of inappropriate content. We check for **curse words** using a check for specific words that are inappropriate, ignoring the context of the words. **Profanity** is focused on language that is disrespectful to religion or god, words like "damn" score high on profanity. However, profanity is broader than religion-related words and also includes a broader sense of inappropriate language, e.g. toxic, obscene, threatening, insulting, hate and offensive speech. **Offensive speech** is speech that is rude or insulting but is not based on race, religion, ethnicity, or other factors that put a person in a vulnerable subgroup. This is the main difference between offensive speech and **hate speech**. Rights for speech [11] describes hate speech as the following: "Hate Speech typically targets the 'other' in societies. This is manifested through the 'othering' of minority groups such as racial, ethnic, religious and cultural minorities, women and the LGBTQI+ community." Quantification of these different kinds of inappropriateness results in a more diverse and detailed overview of inappropriate content.

Curse Words `Google-profanity-words` scans lyrics to check for curse words, using a list of Google's banned curse words[5]. This model allows checking a string to see if it contains any of the words in the list by Google. This check results in a hit or miss result, if the string contains at least one word on the list it results in a hit, otherwise it returns a miss.

`Google-profanity-words` is a widely adopted model to scan for curse words. The Google list is updated monthly to ensure it remains current with the latest developments, it was accessed in May 2024. The model also works quickly, allowing it to handle big datasets such as the large amount of song lyrics.

Profanity The second model, the `Profanity-check` library[6], scans a string to check for profanity. `Profanity-check` is a linear SVM model trained on 200k human-labeled samples of clean and profane text strings. It is the broadest of the four models; scanning profanity in the broad sense of the word, e.g. toxic, obscene, threatening, insulting, hate and offensive speech. Therefore, it has some overlap with the other models. It can result in a binary classification, either profane or not, or it can return a probability of a string being profane. In this experiment, the probability value is used since it gives more information on the profanity level. The following two models also compute a probability value, giving a similar output format.

`Profanity-check` is a widely adopted model to check for profane text in strings. It is not based on a hard-coded list of profane language but on a trained model, allowing it to detect more dynamically if a string is profane e.g. based on context. `Profanity-check` also outperforms a similar dynamic model, `Profanity-filter`, in accuracy and speed performance [24].

Offensive Speech To scan for offensive speech the `Offensive Speech Detector`[7] is used. This model is based on DeBERTa[8], which is a widely adopted language classification model. The offensive speech detector can calculate the probability of a string being offensive.

We use this model since it is trained using a data classification model, allowing it to dynamically scan for patterns to compute the probability of strings being offensive. The performance is acceptable, with an accuracy of 0.747 [2]. Another advantage is that `Offensive Speech Detector` is specifically aimed at offensive speech, causing less overlap with the other models. On the repository of the `Offensive Speech Detector`, the creators argue that it complements the Hate Speech model well in forming a more robust moderation tool.

Hate Speech The last model is similar to the offensive speech model as it is also based on DeBERTa. It is called `Hate Speech Detector`[9]. It detects hate speech in a string and returns the probability of a string being hate speech.

The creators of this model and the `Offensive Speech Detector` model mention each other on their Hugging Face page [1, 2] with the following text: "We believe these models can be used in tandem to support one another and thus build a more robust moderation tool, for example." `Hate Speech Detector` is also aimed at a specific part of inappropriate language, leading to less overlap with other models. This model and the offensive speech detector complement each other well. The last three models mentioned also give a similar output format since they all return a probability value, keeping the format of the results consistent throughout the research.

3.4 Usage of Inappropriate Content Detection Models

Here, we present the usage of the previously presented models for the detection of inappropriate content in the song lyrics.

Binary Presence of Curse Words

To capture the presence of curse words in lyrics, we use the complete lyrics of each song in the database. We scan the lyrics resulting in a hit value or miss value. Either the lyrics contain at least one of the words in Google's list, resulting in a hit; or the lyrics are completely clean of any bad/swear words resulting in a miss.

[5] List of Google's banned curse words: https://github.com/coffee-and-fun/google-profanity-words

[6] Python library profanity-check: https://github.com/vzhou842/profanity-check

[7] Offensive speech detection model: https://huggingface.co/KoalaAI/OffensiveSpeechDetector

[8] DeBERTa language classification model: https://huggingface.co/docs/transformers/model_doc/deberta

[9] Hate speech detection model: https://huggingface.co/KoalaAI/HateSpeechDetector

Percentage of Curse Words

To obtain more information based on the list of bad/curse words from Google, this experiment looks for the percentage of bad/curse words in the lyrics. We achieve this by checking each word in the lyrics of every song to see if it is on Google's list. The amount of curse words in each song divided by the total words in the song times 100, results in the percentage of curse words for that song.

Profanity, Offensive Speech & Hate Speech

The three following experiments are performed similarly. The lyrics of all the songs in the database are input for the models. Then each model computes the probability of the lyrics being profane, offensive or hate speech, corresponding to the model, for all songs per genre.

This results in boxplots with the probability distribution per genre. We compare the distributions based on their mean with a statistical test and based on their spread using the coefficient of variation. This gives an overview of what musical genre is more inappropriate, and the differences between the genres.

4 Results

In this section, we present the results of the experiments conducted to examine the degree to which songs contain inappropriate content.

4.1 Curse Words

We first examine the presence of curse words in the songs, grouped by genre. To quantify the presence of curse words, we rely on the Hit metric and the percentage of curse words.

Based on the Hit metric, we find a percentage of songs per genre that contain at least one curse word. These percentages are shown in Table 3. In general, all genres contain a substantial amount of songs with at least one curse word. Rap stands out with almost nine out of ten songs containing a curse word, R&B also has a majority of songs containing a curse word. The major takeaway of these results is that inappropriate words are present across all five genres.

Genre	Curse	No Curse	%Curse
Pop	1840	3383	35.23%
Rock	895	2425	26.96%
Country	37	127	22.56%
Rap	26434	3616	87.97%
R&B	3005	1532	66.23%

Table 3: Songs containing at least one curse word per genre.

To obtain more information on the amount of inappropriate language in songs based on the list of curse words from Google, we look for the percentage of curse words out of all words in the lyrics. The distribution of these percentages per genre can be found in Figure 1.

The lowest medians in the percentage of curse words in the lyrics are of Pop, Rock, and Country, with values of 0.93%, 1.10% and 0.83% respectively. R&B's mean value is higher at a percentage of 1.51%. The highest percentage of curse words is in Rap music at a value of 2.44%.

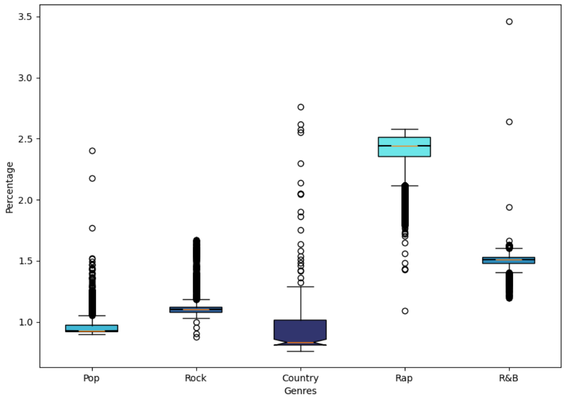

Figure 1: Percentage of curse words in lyrics per genre. Pairwise all differences are significant, except for scores between lyrics in Pop and Country.

The test of homogeneity of variances yields a p-value smaller than $.001$, indicating that the variances across song genres are significantly different. Given this violation of the homogeneity of variances assumption, and the fact that the data distributions are not normally distributed the Kruskall-Wallis test is used to compare distributions. This analysis results in a high H-value (H=28440.6078) and a low p-value ($< .0001$), meaning the distributions differ significantly between genres.

Post Hoc tests with Dunn's test, which is not affected by outliers and can be used for multiple means, show pairwise significant differences between all genres, except between Pop and Country music. In this post hoc test, Bonferroni correction is applied to prevent false positives in a test with multiple comparisons. For eight out of ten pairs, the p-value is smaller than $.0001$, indicating very high significance. For Rock and Country, the p-value is between $.01 < p <= .05$, indicating significance but less convincingly. The difference between Pop and Country is not significant.

Noticeable in the boxplots is the small spread of curse words in songs of the same genre, visualised by the small interquartile range and whiskers in the graph. The coefficient of variation (CV) represents how far apart the data points differ from the mean in a distribution. If the CV is smaller than 1, the variability of data is considered small. In this case, the CV scores range from 0.0437 (R&B) to 0.3935 (Country). This means that the data points are very close to the mean, thus many songs within a genre have a similar percentage of curse words. The small whiskers however do cause many outliers across all genres.

Overall, Rap music is the most inappropriate, with a significantly higher percentage of curse words than all other genres, with almost twice as many curse words as R&B per song and over twice as many as the other three genres. R&B stands in second place and Rock in third place, both with significant differences from all other genres. Pop and Country both have the smallest percentage of curse words per song. Per genre songs have similar percentages of curse words.

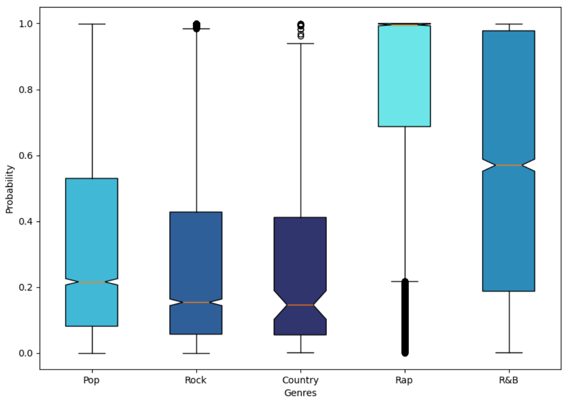

Figure 2: Probability distribution of songs to be profane per genre. Pairwise all distributions differ significantly, except (Pop, Country) and (Rock, Country).

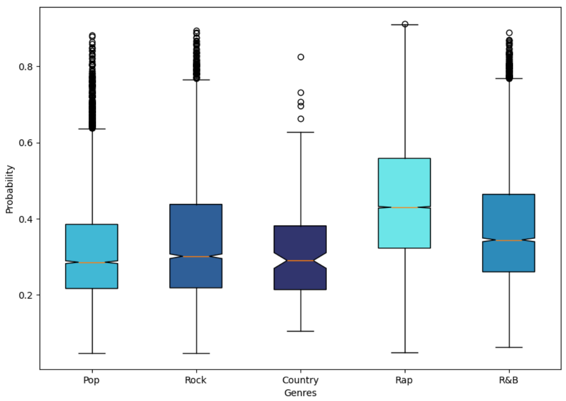

Figure 3: Probability distribution of songs to be offensive per genre. Pairwise all distributions differ significantly, except (Pop, Country) and (Rock, Country).

4.2 Profanity

This experiment measures the presence of profanity in the music lyrics of each genre. This presence is visualized in Figure 2.

The median of Rap is at a probability value of 0.9955, higher than the other genres. A median value of 0.9955 means that half of all songs have a probability of at least 99.55% being profane, meaning it is almost certain to have profanity. R&B stands in second place in profanity probability, with a median of 0.5696. Pop, Rock, and Country have the lowest medians at 0.2156, 0.1533, and 0.1454 respectively.

To measure significance a similar test is performed as in section 4.1. Kruskal-Wallis test gives an h-value of H=10843.3367 and a p-value of $p < .0001$, meaning there is a high significant difference between genres.

Dunn's test, with Bonferroni correction, shows pairwise significant differences between all genre pairs, except for the two pairs (Pop, Country) and (Rock, Country). For all eight pairs of genres with significant differences, the p-value is $< .0001$, indicating high significance.

The spread of these graphs is larger than the curse word result. The CV scores of the genres range from 0.4057 (Rap) to 1.076 (Country). This means the data points are not especially close to the mean value but are also not spread out far, thus the probability of a song being profane can vary slightly within a genre.

For the two pairs with insignificant differences, we cannot conclude on their inappropriateness level. For the other genres, it can be said that Rap has the significantly highest profanity probability, meaning songs in Rap are the most likely to be profane. R&B stands at a clear second place, significantly more likely to be profane than Pop, Rock, and Country. Pop is also significantly more likely to contain profanity than Rock music.

4.3 Offensive Speech

The result of the experiment measuring offensive speech can be found in Figure 3.

A result similar to the previous 2 models is achieved. Rap music again has the highest mean (0.4434) and median (0.4300) value. R&B is in second place with a median of 0.3445. Pop, Rock, and Country have lower medians at 0.2859, 0.3015, and 0.2903 respectively. The results from this model are slightly more nuanced however, all distributions have a larger overlap than the curse words result and the medians are closer together than in the previous results.

Similarly to the previous result analysis, a Kruskal-Wallis test is used to check if the difference in distributions is significant. This results in a test statistic of H=3942.5339 and a p-value of $p < .0001$. The post hoc test with Dunn's test gives similar results to the profanity results (Section 4.2). The difference between Pop and Country; and Rock and Country is not significant, meaning we can not say anything about those differences. All other pairs of genres have a p-value of $p < .0001$, indicating very high significance.

The interquartile range and size of the whiskers are more similar to the results of the offensive speech model. This is also signified by the similar CV scores across genres, ranging from 0.3775 to 0.4951. This small range in CV scores mostly shows that this model creates similar distributions for each genre.

From this information, we can conclude that Rap again scores highest in inappropriateness. R&B takes second place again. Then follow Pop, Rock, and Country, where Rock scores higher than Pop in the amount of offensive speech. The distributions as results of the offensiveness model are similar to previous experiments.

4.4 Hate Speech

The final experiment measures the probability of a song being hate speech. The results of this experiment can be found in Figure 4.

The results from this experiment differ from the previous experiments. Rap is no longer the most likely to be inappropriate. Rock is the most likely to be hate speech with a mean probability of 0.4610. The medians of Pop, Country, Rap,

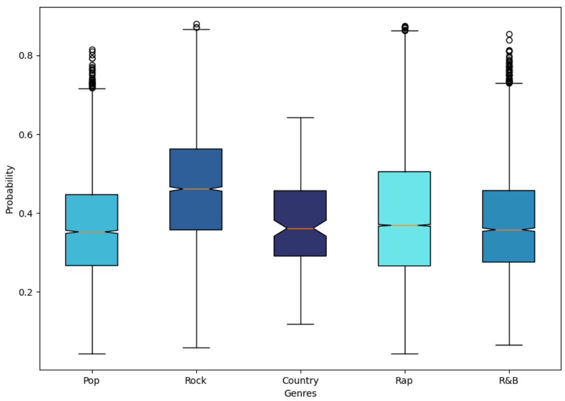

Figure 4: Probability distribution of songs to be hate speech per genre.

and R&B are closer together at 0.3522, 0.3616, 0.3688, and 0.3578 respectively.

We measure significance with a Kruskal-Wallis test, resulting in a test statistic of H=937.8618 and a p-value of $p < .0001$. Dunn's test gives the pairwise significance, it shows hate speech in Rock music is significantly more probable than in all other genres with a p-value of $p < .0001$. Country music has no significant differences with Pop, Rap, and R&B, so we cannot conclude anything based on their differences. Rap scores significantly higher than Pop and R&B, with a p-value of $p < .0001$. Finally, R&B also scores higher on hate speech than Pop music, with a p-value of $.01 < p <= .05$, still indicating significance but less convincingly.

The distributions of songs resulting from the hate speech model are similar to the offensive speech model. The CV score of the hate speech distributions ranges from 0.3110 (Country) to 0.4329 (Rap), which means the distributions are similar in variability. The relatively low CV score (> 1 is considered large variability), indicates that within each genre the songs are close to the mean value, and thus have similar probabilities of being hate speech.

The scan for hate speech in song lyrics resulted in different results than the previous models. Rock is labelled as the most inappropriate, based on hate speech. Rap also contains more hate speech than Pop and R&B. The distributions per genre of hate speech are more similar to each other than the distributions of curse words and profanity.

5 Discussion

Here, we discuss the explanation and implications of the results and place the results in the context of children's music recommender systems.

Some songs in the database are not completely English; some are a combination of foreign languages and English, and some are mostly foreign with some English words throughout the lyrics. Since this research is aimed at an English-speaking audience, the models only recognize inappropriate content in English. Therefore, if a song contains many foreign lyrics the models might not classify appropriate or inappropriate clearly. However, the words recognizable by English speakers and thus possibly inappropriate to English speakers, are recognized by the models and thus labelled as inappropriate when applicable.

According to the models used in this work, Rap scores highest on inappropriateness based on the percentage of curse words, the probability of containing profanity, and the probability of containing offensive speech. The hate speech model ranked Rap in second place, significantly higher than Pop and R&B. It can be concluded that Rap jumps out as the most inappropriate musical genre, it is therefore likely not suitable for children. The high percentage of curse words per song (2.44% for Rap, 1.51% for the second highest, R&B) can cause children to learn inappropriate language. Children tend to repeat words they often hear, and with lyrics where over 2% of all words are curse words, it seems inevitable that children's speech behaviour will be negatively influenced. Recommender systems should therefore be cautious with recommending Rap music to children.

Similar results to Rap are found for R&B. R&B is placed second in inappropriateness by all models that label Rap as most inappropriate. The hate speech model ranks R&B behind Rock and Rap, but it scores significantly higher than Pop. These high scores for R&B can partially be caused by the large overlap in songs between R&B and Rap. Over 50% of songs in R&B are also in Rap as shown in Table 2. The inappropriateness level, therefore, is similar, as shown in the results. Similarly to Rap music, recommender systems should be cautious with recommending R&B music to children.

The offensive speech model and the hate speech model are both trained on the same dataset and both use the same basis model (deBERTa). The distributions they output therefore look similar, with CV scores of the offensive speech model ranging from 0.3775 to 0.4951 and the CV scores of the hate speech model ranging from 0.3110 to 0.4329. The median values are also close together, 0.2859 to 0.4300 for offensive speech and 0.3522 to 0.4610 for hate speech. However, despite the similar distributions, the results differ. While Rock scores relatively low in inappropriateness for the curse word, profanity and offensive speech models, it scores highest in hate speech. This could indicate that Rock contains a less obvious and more specific form of inappropriate content in the form of hate speech. This shows that inappropriateness is not a binary term that can label every song as appropriate or inappropriate, but it is more diverse. This indicates that binary labels like "Explicit" by Spotify or "parental advisory" for the general music industry do not expose the more nuanced forms of inappropriateness. Recommender systems should therefore not base their recommendations on a binary label of (in)appropriateness; a more nuanced classification is advised to cover all possibly harmful forms of inappropriate content.

Profanity, offensive speech and hate speech, in contrast to curse words, are not always as clearly recognisable since the models base their probabilities on context. The genres with high scores in these categories (Rap, R&B, and Rock) might not directly teach children inappropriate words or sentences. However, exposing children to profane text, offensive speech, and hate speech could also lead to less directly noticeable behavioural and personality changes, for example, by creating a bias against certain ethnicities, religions, or social groups. This underscores the importance of nuanced content filtering to protect children from these indirect but potentially harmful effects.

We advise a nuanced filter to filter inappropriate lyrics. This filter should consider different facets of inappropriateness per song and tailor the recommendations to the needs of children. This research has covered several of these facets and shows clear differences between genres. However, we do not claim to have covered the complete overview of inappropriate content. The models used in this paper focused on curse words, profanity, offensive speech, and hate speech, but future work could expand on this. Examples of inappropriate content we did not cover explicitly in this paper are sexual content, misogyny, violence, and drug/alcohol use. Some of these categories fall under this research's models, but explicit research could improve the in-depth overview of all forms of inappropriate content to give more informed advice to future recommender systems.

From all significant results, it can be concluded that Pop is the least inappropriate. This genre thus seems the best fit to recommend to children, but it is still not completely child-friendly. This is shown by the Hit property of the curse word scan that shows that over a third of songs contain at least one inappropriate word (Table 3). This suggests that while Pop is relatively safer, it still requires careful screening to ensure suitability per song for children.

Throughout the results, it is noticeable that Country music has few significant results in comparison to the other genres. This is caused by the relatively small dataset of 164 Country songs, compared to 3320 songs in the second smallest genre, Rock (see Table 1). Therefore, we cannot conclude much about inappropriate lyrics in Country music. From the significant comparisons, we can conclude that Country is less inappropriate than Rap and R&B.

To put our findings in context, we compute the readability levels of the song lyrics to understand the possible impact of inappropriate content. Readability levels associate a sample text with a grade level indicating the target audience that could comprehend the text. Table 4 shows what age relates to what grade. To get the readability levels we use three different methods; Flesh-Kincaid, because of its wide adoption; Spache, since it is not only based on sentence length but also familiar words and it works well for texts aimed at children up to fourth grade [22]; and Dale-Chall, because it is also based on familiar words and works well across a wider variety of age ranges, compensating for the grade levels that Spache's method might underperform [8].

To achieve usable results, we added punctuation to the lyrics by replacing new-line characters (“\n”) with a period (“. ”). This is necessary since all three formulas use sentence length to compute readability. This substitution method might not result in perfect punctuation. This is why we chose two formulas that also rely on familiar words to compensate for the punctuation.

Figure 5 shows the readability scores measured by the Flesch-Kincaid, Spache, and Dale-Chall formulas. The readability tests overall clearly show that young children can understand song lyrics. The highest mean value results from Dale-Chall for Rap music, at 7.9855, which relates to an age range of 13-14. Spache results in means ranging from 3.5740 to 4.6561, which relates to children of 8-11. Means by Flesch-Kincaid range from 0.6685 to 1.6945, indicating children of 5-8. This shows children of young ages can understand the song lyrics and can thus be influenced by the lyrics, reaffirming the need for recommender systems to be cautious in their recommendations to children.

School Grade	Age	School Grade	Age
Preschool	4-5	Grade 6	11-12
Kindergarten	5-6	Grade 7	12-13
Grade 1	6-7	Grade 8	13-14
Grade 2	7-8	Grade 9	14-15
Grade 3	8-9	Grade 10	15-16
Grade 4	9-10	Grade 11	16-17
Grade 5	10-11	Grade 12	17-18

Table 4: Grade and age correspondence.

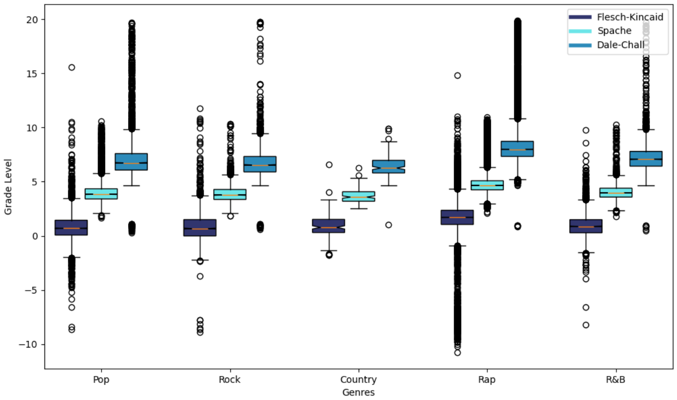

Figure 5: Flesch-Kincaid, Spache, and Dale-Chall readability scores per genre. Six outliers scoring higher than grade level twenty, four in Rap and two in Rock, were removed to improve the readability of the graph.

All three formulas label Rap as the most difficult to understand. R&B ranks second in all tests. So a similar trend is noticeable in readability scores compared to inappropriateness level. This means that the inappropriate genres are more difficult to understand to children. This could mediate the risk of inappropriate genres negatively influencing children since children do not understand the lyrics. However, from fourteen years old and onwards all three formulas say that children on average can understand the lyrics, which is still a highly developmental age, vulnerable to negative influences.

Overall, the results show that inappropriate content is present in music lyrics of Pop, Rock, Country, Rap, and R&B. The Hit property has shown that a substantial portion of songs per genre contain at least one inappropriate word. The other models have shown that different forms of inappropriateness are present in music and in what proportions they are present. Recommender systems should be cautious when recommending Rap and R&B music to children since they are the most inappropriate. Pop music seems the best fit to recommend to children. However, filtering on genres seems drastic and unnuanced, resulting in possibly filtering acceptable songs from Rap and R&B, or letting inappropriate songs from Pop slip through. We also show that children can understand the music lyrics from a young age and thus be influenced by them. Therefore, we advise children's music recommender systems to include a filter on lyrics to check if the song is inappropriate and in what way it is inappropriate, before recommending a song to children, to avoid negatively influencing them.

6 Responsible Research

This research uses a dataset and models to scan for inappropriate content. Using such data comes with the responsibility to handle data ethically.

Children The results of this paper aim to guide future recommender systems to prevent recommending inappropriate content to youngsters. Children are a vulnerable target audience, so extra precautions should be taken to mitigate the risk of badly influencing them. Research with children as the main focus should consider these precautions at all times. This research does not handle children's data directly, making the possibility of misusing their data unlikely. The outcome of this research aims to protect children from inappropriate music lyrics, aligning with the broader goal of promoting a safe and positive online environment for children.

Data Data in the Genius database can be added by users of the Genius platform and is publicly available for free. This data contains no personal information besides users' usernames who add songs or song annotations. These usernames are not used for this research. All use of data should comply with ethical standards and legal regulations regarding data privacy.

Verification of Results The Genius database and the 4 models to scan inappropriate content are also publicly available for free. Anyone can reenact and verify all results obtained from using the models on the database. All preprocessing steps have been described in this paper, this enables the reproducibility of this research. All code used to obtain results in this paper can be found in the repository mentioned in Section 3.

Black Box Models The models described in this paper are trained on databases, these databases can contain biases that the models enforce unintentionally. Biases in data or algorithms can lead to unfair outcomes, such as disproportionately labelling certain genres as inappropriate. The databases that the models are trained on are publicly available, so the data that is used is known. However, detecting biases in this data and in the models is difficult. This difficulty arises because biases are often subtle, and embedded in the data in ways that are not immediately apparent. Standard validation methods may not reveal these biases since both the training and validation data of the model share the same flaws.

The models, which are trained on a dataset, do not give reasoning for the values they output. The models can make mistakes, or base their output on biased data. However, due to a lack of reasoning in the output, these mistakes are not noticeable. However, the way that the models are trained and the data the models are trained on, are publicly available. This allows anyone to retrain the models making the results more reproducible. This research also considers multiple models to mitigate the possible bias one model could have by scanning for different facets.

7 Conclusions and Future Work

In this work, we aimed to create an in-depth overview of inappropriate content in Pop, Rock, Country, Rap, and R&B music. This was shown by an empirical exploration of songs in the Genius database using four facets of inappropriateness: curse words, profanity, offensive speech and hate speech. The findings of our empirical exploration show inappropriate content is present in music across all five genres and children from an early age can understand the lyrics within these genres. This shows that present music can be inappropriate to children and therefore recommender systems should be cautious when recommending songs to children.

Results show that Rap and R&B are the most inappropriate genres, possibly teaching children curse words, profane language, offensive speech, or hate speech. Recommender systems could have a filter for these genres, but filtering on entire genres is drastic. Therefore, we advise future children's music recommender systems to consider a filter per song to check if it is inappropriate to recommend to children and in what way it is inappropriate. This is also well suited for Rock music, which is likely to contain hate speech, but scores low in the other facets of inappropriateness. A filter recognising this hate speech could tailor its recommendations to suit children.

The results and conclusions of this paper are purely based on the four selected methods to scan inappropriateness. These four models do not show the complete landscape of inappropriate content. Future research can expand the results of this paper by using different models to scan for other facets of inappropriate content.

The genres in this paper are logically selected from the Genius dataset. However, they do not show the entire musical landscape youngsters listen to. Using the secondary genre tags in the Genius database or using a different database in future research can expand the musical broadness of this research.

In conclusion, we advise future recommender systems to be cautious with recommending music to youngsters. Especially the genres Rap and R&B are likely to be inappropriate. This study underscores the need for enhanced content filtering in music recommender systems to protect young listeners from inappropriate content.

References

[1] Koalaai/hatespeechdetector · hugging face, 01 2024.

[2] Koalaai/offensivespeechdetector · hugging face, 01 2024.

[3] SimilarWeb - Genius.com, 4 2024.

[4] Terri M. Adams and Douglas B. Fuller. The words have changed but the ideology remains the same: Misogynistic lyrics in rap music. *Journal of Black Studies*, 36(6):938–957, 07 2006.

[5] Susan L Andersen. Trajectories of brain development: point of vulnerability or window of opportunity? *Neuroscience & Biobehavioral Reviews*, 27(1-2):3–18, 01 2003.

[6] Jeffrey Jensen Arnett. Heavy metal music and reckless behavior among adolescents. *Journal of youth and adolescence*, 20(6):573–592, 12 1991.

[7] Albert Bandura and Richard H Walters. *Social learning theory*, volume 1. Englewood cliffs Prentice Hall, 1977.

[8] Jeanne Sternlicht Chall and Edgar Dale. *Readability Revisited : The New Dale-Chall Readability Formula.* Brookline Books, 01 1995.

[9] Ying Chen, Yilu Zhou, Sencun Zhu, and Heng Xu. Detecting offensive language in social media to protect adolescent online safety. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, pages 71–80, 2012.

[10] Hyojin Chin, Jayong Kim, Yoonjong Kim, Jinseop Shin, and Mun. Y. Yi. Explicit content detection in music lyrics using machine learning. In *2018 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pages 517–521, 2018.

[11] Rights for Peace. What is hate speech?

[12] Cynthia M. Frisby and Elizabeth Behm-Morawitz. Undressing the words: Prevalence of profanity, misogyny, violence, and gender role references in popular music from 2006-2016. *Media watch*, 10(1), 01 2019.

[13] Denise Herd. Changing images of violence in rap music lyrics: 1979–1997. *Journal of Public Health Policy*, 30(4):395–406, 12 2009.

[14] Kyle J. Holody, Christina Anderson, Clay Craig, and Mark Flynn. “drunk in love”: The portrayal of risk behavior in music lyrics. *Journal of Health Communication*, 21(10):1098–1106, 09 2016.

[15] L. Rowell Huesmann, Jessica Moise-Titus, Cheryl-Lynn Podolski, and Leonard D. Eron. Longitudinal relations between children’s exposure to tv violence and their aggressive and violent behavior in young adulthood: 1977-1992. *Developmental Psychology*, 39(2):201–221, 2003.

[16] Ashawnta Jackson. Parental advisory: The story of a warning label, 09 2020.

[17] Graham Martin, Michael Clarke, and Colby Pearce. Adolescent suicide: Music preference as an indicator of vulnerability. *Journal of the American Academy of Child & Adolescent Psychiatry*, 32(3):530–535, 05 1993.

[18] Kostantinos Papadamou, Antonis Papasavva, Savvas Zannettou, Jeremy Blackburn, Nicolas Kourtellis, Ilias Leontiadis, Gianluca Stringhini, and Michael Sirivianos. Disturbed youtube for kids: Characterizing and detecting inappropriate videos targeting young children. *Proceedings of the International AAAI Conference on Web and Social Media*, 14(1):522–533, 05 2020.

[19] Amifa Raj, Ashlee Milton, and Michael D. Ekstrand. Pink for princesses, blue for superheroes: The need to examine gender stereotypes in kid’s products in search and recommendations, 05 2021.

[20] Elaine Scharfe. Development of emotional expression, understanding, and regulation in infants and young children. 01 2000.

[21] Md Abdus Salam Siddique, Md Imran Sarker, Robin Ghosh, and Kamal Gosh. Toxicity classification on music lyrics using machine learning algorithms. In *2021 24th International Conference on Computer and Information Technology (ICCIT)*, pages 1–5, 2021.

[22] George Spache. A new readability formula for primary-grade reading materials. *The Elementary School Journal*, 53(7):410–413, 03 1953.

[23] Joe Whittaker, Seán Looney, Alastair Reed, and Fabio Votta. Recommender systems and the amplification of extremist content. *Internet Policy Review*, 10(2), 06 2021.

[24] Victor Zhou. Building a Better Profanity Detection Library with scikit-learn, 2 2019.