AUTOMATIC DETECTION OF MORAL VALUES IN MUSIC LYRICS

Vjosa Preniqi[1] **Iacopo Ghinassi**[1] **Julia Ive**[1]
Kyriaki Kalimeri[2] **Charalampos Saitis**[1]
[1] Centre for Digital Music, Queen Mary University of London, London, UK
[2] ISI Foundation, Turin, Italy
{v.preniqi, i.ghinassi, j.ive, c.saitis}@qmul.ac.uk, kyriaki.kalimeri@isi.it

ABSTRACT

Moral values play a fundamental role in how we evaluate information, make decisions, and form judgements around important social issues. The possibility to extract morality rapidly from lyrics enables a deeper understanding of our music-listening behaviours. Building on the Moral Foundations Theory (MFT), we tasked a set of transformer-based language models (BERT) fine-tuned on 2,721 synthetic lyrics generated by a large language model (GPT-4) to detect moral values in 200 real music lyrics annotated by two experts. We evaluate their predictive capabilities against a series of baselines including out-of-domain (BERT fine-tuned on MFT-annotated social media texts) and zero-shot (GPT-4) classification. The proposed models yielded the best accuracy across experiments, with an average F1 weighted score of 0.8. This performance is, on average, 5% higher than out-of-domain and zero-shot models. When examining precision in binary classification, the proposed models perform on average 12% higher than the baselines. Our approach contributes to annotation-free and effective lyrics morality learning, and provides useful insights into the knowledge distillation of LLMs regarding moral expression in music, and the potential impact of these technologies on the creative industries and musical culture.

1. INTRODUCTION

Lyrics play a crucial role in how we experience music, affecting our emotions and actions. Positive lyrics can motivate and elevate listeners, whereas negative or aggressive content in songs may negatively impact mood and behaviour [1]. Social, political, and cultural issues, such as racial inequality and gender discrimination, are often reflected in the music lyrics of their time [2, 3]. Songs that feature in successful campaigns typically include uplifting melodies and lyrics that reflect the ideals of a nation, representing values of optimism and progress towards a better future [4]. Moral rhetoric in lyrics has been used to advocate for what is perceived to be a necessary societal change [5], promote peace and unity [6], and raise awareness for marginalised groups [7]. These narratives are closely related to moral judgements and beliefs, yet their relationship to music listening behaviors has received limited attention by music scientists.

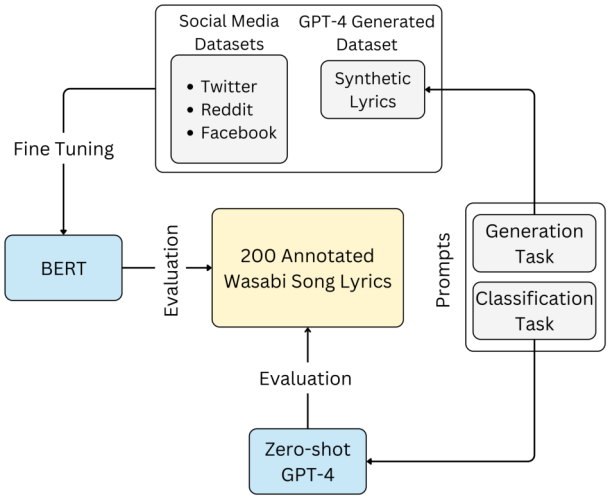

Figure 1. Model Structure for predicting Moral Foundations (MFT) in Lyrics, fine-tuned on out-of-domain social media data, and synthetically generated lyrics with GPT-4.

In the field of Music Information Retrieval (MIR), lyrical content analysis has focused primarily on genre classification [8], mood prediction [9], emotion dynamics [10], and lyrics-to-audio alignment [11, 12]. Recent works have elaborated on less attended psychological characteristics of music lyrics, including moral valence. For example, insights into personal values and personality traits derived from lyrics can enhance various MIR tasks, including genre classification, audio tagging, and music recommendations [13]. Preniqi and colleagues [14] showed that moral valence extracted from lyrics can to some extent predict listeners' moral values, in some cases more accurately than audio features. The possibility to extract morality rapidly from lyrics can enable a deeper understanding of our music listening behaviours.

Inferring moral values from song lyrics is a complex natural language processing (NLP) task from the start due to the subjectivity of our perceptions and interpretations. The progress is further hindered by the lack of an-

notated lyrics for training new or fine-tuning pre-trained models, and for benchmarking. Using models fine-tuned with out-of-domain annotated texts (e.g., from social media [15, 16]) to predict moral values in music lyrics faces significant challenges due to the unique structure of lyrics compared to other textual forms (e.g., greater use of repetition, metaphor, imagery, and other poetic devices).

In light of the above, we investigate the novel task of automatic detection of moral values in music lyrics using an integrated approach that leverages the strengths of two distinct NLP technologies. Specifically, we leverage the generative capabilities of GPT-4 (Generative Pre-trained Transformer) to create morally nuanced synthetic lyrics—a process required only once—and employ BERT (Bidirectional Encoder Representations from Transformers), which demands fewer computational resources, to learn from the synthetic data structure.

Following recent related work [14, 15, 17], we operationalize morality drawing on Haidt and Graham's Moral Foundations Theory [18], which outlines five core moral traits, or foundations, divided into "virtue" and "vice" based on moral polarity: *Care* and *Harm*, *Fairness* and *Cheating*, *Loyalty* and *Betrayal*, *Authority* and *Subversion*, *Purity* and *Degradation*. We developed a corresponding set of 10 single-label classification models, each customized to predict the presence or absence of one moral value in lyrical text. MFT is a straightforward yet comprehensive model for understanding moral values, uniquely characterized by well-developed term dictionaries [19].

We present a dataset of 200 real song lyrics human-annotated with MFT. To the best of our knowledge, this is the first such dataset. It serves as the basis for evaluating our proposed method. We make the real and synthetic lyrics datasets, and the paper code fully available via a GitHub repository. [1]

We report a comprehensive comparison of the proposed models against BERT fine-tuned with out-of-domain human-annotated moral text data and zero-shot classification with GPT-4. Figure 1 summaries the overall pipeline of this work. The proposed models yielded the best accuracy across experiments, with an average F1 weighted score of 0.8. This performance is, on average, 5% higher than out-of-domain and zero-shot models. When examining precision in binary classification, the proposed models perform on average 12% higher than the baselines. Our approach contributes to annotation-free lyrics morality learning, and provides useful insights into the knowledge distillation of large language models such as GPT-4 regarding moral expression in music.

2. RELATED WORK

The field of music and moral expression has received limited attention. However, recent studies have shown a link between an individual's moral values and their preferences for lyrics and music, suggesting significant implications for tailoring personalisation in streaming services [14, 17, 20]. Further research has delved into how moral values and lyrical preferences manifest within specific music communities. For example, Messick and Aranda [21] demonstrated that moral values could explain a unique and significant portion of the variance in lyrical preferences among fans of different metal music sub-genres.

Given the understanding that verbal expressions more effectively convey morality than non-verbal forms [17,22], initial studies introduced lexicons [23, 24] as an extension of Moral Foundations Dictionary (MFD) [25] for identifying words and lemmas that accurately depict moral foundations. More recent studies focused on examining moral values in texts using human-annotated social media datasets [26–28], and introducing more advanced Natural Language Processing (NLP) approaches to detect moral dimensions in textual content [15, 16]. Trager et al. [27] introduced baseline models for predicting moral values, employing a pre-trained BERT model fine-tuned on the Moral Foundation Reddit Corpus. Guo et al. [16] proposed a multi-label model for predicting moral values with Twitter and news data, incorporating the domain adversarial training framework suggested by Ganin et al. [29] to align multiple datasets and generalise for out-of-domain predictions. A similar approach was taken by Preniqi et al. [15] in predicting moral values in different social media domains.

However, a main challenge that persists is the ability of these models to generalise across various domains. Lisco and colleagues [30] demonstrated that text classifiers perform better when domains are similar. This poses a major obstacle when predicting morality in lyrics because there is no prior study that has presented an annotated lyrics dataset with moral values. Further, manually annotating extensive text demands substantial time, resources, and deep understanding of Moral Foundations Theory (MFT).

To overcome these limitations, we employ GPT-4, an advanced LLM, to generate lyrics infused with various moral undertones, which helps in fine-tuning a moral classifier. This minimises the need for laborious manual annotation of extensive lyric databases, enabling us to utilise a smaller, human-annotated dataset to validate the effectiveness of knowledge distilled from GPT-4. The capacities of LLMs for music tasks are being actively explored for the moment. Doh et al. [31] similarly employed a large language model such as GPT-3 for generating pseudo captions from tags to mitigate the problem of data scarcity in the field of automatic music captioning. While Zhang et al [32] evaluated the quality and correctness of generated music lyrics via GPT-3. Sawicki et al. [33] investigated the possibility of using GPT-3 models to generate high-quality poems in a specific author's style while suggesting that GPT-3 can be a useful tool in assisting authors.

3. METHOD

3.1 Human-Annotated Lyrics

For this work, we annotated 200 song lyrics, categorising them into 10 different moral foundations. This annotation process was conducted by two skilled annotators: the

[1] https://github.com/vjosapreniqi/ismir-mft-values

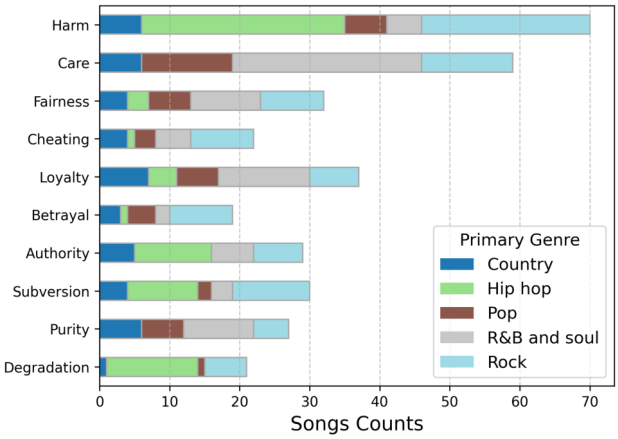

Figure 2. Distribution of Moral Foundations in 200 song lyrics dataset annotated by human annotators with genre proportions for each moral foundation.

lead author of this study and an external researcher with a background in music and sound design, both of whom agreed to contribute. Before starting, the annotators were informed about their participation rights, including the option to discontinue their involvement at any point. Each annotator was assigned with 125 songs for annotation. To evaluate the agreement between annotators, 50 songs were annotated by both annotators. The inner-annotator agreement was assessed using Cohen's kappa coefficient for each moral label. This resulted in an almost perfect agreement [34] with an average score of 0.86 across all moral categories identified within the lyrics of the chosen songs. We selected the songs for the moral values annotation from the Wasabi Dataset [35], known for its extensive collection of 2 million songs including lyrics, artist gender, and musical genre among other data. This dataset spans over five decades, enabling the selection of songs from various eras. The process of selecting the songs involved a semi-random approach, with efforts made to retain the distribution of genres, and the timeline of song releases as found in the original dataset. Among the 200 songs annotated for moral values, 18 were from the 60s and 70s, 78 from the 80s and 90s, and 116 from the post-2000 era. The chosen songs represented a balanced mix of genres including Rock, Pop, Hip-Hop, R&B, Soul, and Country. Figure 2 depict the distribution of Moral values in the human-annotated song lyrics with the proportion of genre for each MFT value.

3.2 Predicting Morality in Lyrics with Domain Adaptation

Initially, we tried to predict moral values in lyrics by fine-tuning a BERT model with out-of-domain social media data, following the approach used by Preniqi et al. [15]. We utilised 20,628 tweets from the Moral Foundation Twitter Corpus (MFTC) [26]; 13,995 posts from the Moral Foundations Reddit Corpus (MFRC) [27]; 1,510 posts from Facebook vaccination dataset [28]. Preniqi's and other work have demonstrated that predicting moral values using a single-label approach—predicting one MFT value at a time—results in higher accuracy [15, 27]. Informed by these findings, we developed a set of single-label classification models tailored to predict individual moral foundations in lyrics.

As a baseline model, we apply a similar approach to the MoralBERT [15]. We identify the polarities (virtues and vices) of moral foundations, as opposed to just identifying the mere presence or absence of moral values. We incorporate the domain adversarial method aiming to improve the models' ability to generalise effectively in predicting moral values in lyrics [15, 16]. Adopting this model, we start by deriving a domain invariant representation h from the BERT CLS embedding e:

$$h = W_{inv}e$$

where $W_{inv} \in \mathcal{R}^{768\times768}$ is a learnable matrix. Next, we calculate moral values predictions $\hat{y_m}$ using:

$$\hat{y_m} = Softmax(W_1(ReLU(W_2h)))$$

with $W_1 \in \mathcal{R}^{768\times768}$, $W_2 \in \mathcal{R}^{768\times c}$ representing 2 learnable matrices, c being the number of classes, $ReLU$ is the rectified linear unit activation function and $Softmax$ is the normalised exponential function. A domain classification head is also included for obtaining domain predictions $\hat{y_d}$:

$$\hat{y_d} = Softmax(W_3(ReLU(W_4h)))$$

with $W_3 \in \mathcal{R}^{768\times768}$, and $W_4 \in \mathcal{R}^{768\times d}$ learnable matrices and with d being the number of domains in the training set. The main rationale of the adversarial network is increasing the loss from the domain head while minimising the loss from the moral values prediction. Hence, the model is "forced" to learn domain-invariant representations. This is achieved by integrating a gradient reversal layer before the domain classification head, while using standard training for minimising moral prediction loss. Cross-entropy (CE) loss is used for both the moral and domain classification heads. The final loss is expressed as:

$$L = CE(\hat{y_m}, Y_m) - CE(\hat{y_d}, Y_d) + L_{norm} + L_{rec}$$

with and Y_m and Y_d as the ground truth for moral values and domain, respectively. Two regularisation terms from [16] are added: L2 norm regularisation and reconstruction loss:

$$L_{norm} = ||W_{inv}h - I||^2, \quad L_{rec} = ||W_{rec}h - e||^2$$

similar to W_{inv} (defined above), $W_{rec} \in \mathcal{R}^{768\times768}$ is also a learnable matrix and I is the identity matrix. These regularization losses are combined with moral and domain classification losses. The regularization terms are not applied when training MoralBERT on a single domain (e.g., when trained on just synthetic lyrics).

The binary setting we use implies the model should learn from highly unbalanced datasets, where the neutral label (negative class) is far more represented than the single moral value to be predicted in each instance (positive

class). To address the class imbalance, we employed two methods. First, weights are assigned to classes [36]:

$$weight_c = \frac{N - N_c}{N}$$

where N is the total training samples and N_c is the count of samples per class c. Second, similar to [37], we employed a separate threshold θ_v for each moral value v, so that we use $\hat{y_m}$ to obtain the final prediction $\hat{m}$:

$$\hat{m} = \begin{cases} 1 & \text{if } \hat{y_m} > \theta_v \\ 0 & \text{otherwise} \end{cases}$$

with $\hat{m} = 1$ indicating the moral value is present in the lyrics and $\hat{m} = 0$ indicating it is not. The optimal value θ_v for each moral value v was found by optimizing for binary F1 during training, searching in the search space 0.05 to 0.95 with a step of 0.05. The models were trained for 20 epochs using a single Nvidia T4 GPU, a learning rate of 5e-5, and the Adam optimiser for all MoralBERT experiments.

3.3 Synthetic Lyrics Generation for Moral Assessment

There is a growing interest in knowledge distillation from large pre-trained language models via synthetic text generation [38]. Here we apply a similar knowledge distillation approach by utilising GPT-4 for synthetic lyrics generation. This method eliminates the need to collect real-life data, which is often difficult to gather for a specific NLP task and with a specific input distribution [39]. Initially, we assessed GPT-4's familiarity with Moral Foundation Theory [25], confirming its fundamental understanding of moral values. We tasked GPT-4 with generating lyrics by formulating a prompt, as follows:

Prompt: *You are an assistant to a songwriter, you need to assist in writing lyrics related to the Moral foundations described in the Moral Foundation Theory. Given the* {`Moral Foundations Tags`} *, which represent* {`Description Tags`}*, write original lyrics of a song expressing these moral foundations. DO NOT directly mention these moral foundations. DO NOT explicitly talk about morality. Write it in the style of* {`Artist Tags`}.

We assigned a "role" (songwriter assistant) for the model and provided three types of "input tags". The {`Moral Foundations Tags`} comprise any of the 10 moral values. The resulting lyrics can represent 1, 2, or 3 moral values. We determined this based on the moral combinations observed in our human-annotated lyrics dataset. The {`Description Tags`} represent fundamental concepts of each moral value. The {`Artist Tags`} represent the names of artists whose styles we employ to diversify the lyrics. Initially, we intended to commence the lyrics generation task solely using moral categories and genres as tags. However, we observed that the lyrics were more uniform and generic compared to when we incorporated the artist's style. To tailor the lyrical style using various artists, we employed MusicOSet [40], a collection of musical elements (e.g., music, albums, artists, genres and popularity) suitable for music data mining. To capture the nuances of different genres, we organized the artists according to their popularity and grouped them into prevalent genres like Rock, Pop, Country, Hip Hop, R&B, Soul, Folk, Blues, and Jazz. These genres align very closely with those in the song lyrics we selected for human annotations. We chose to utilise this dataset because it offers detailed data on artist genres and sub-genres, as well as an artist popularity metric that we employ in developing lyric styles. We acquired a dataset comprising 2,721 artificially generated lyrics, each aligned with moral categories similar to our human-annotated lyrics dataset. On average, the generated lyrics had 146 words, with a total of 10,305 unique words across the synthetic lyrics dataset.

3.4 GPT-4 in Moral Classification Task

In addition, we wanted to assess the capability of the 0-shot GPT-4 model in classifying morality in actual song lyrics while comparing it to our proposed model. To do so, we prompted the task as follows:

Prompt: *You will be provided with song lyrics. The song lyrics will be delimited with #### characters. Classify each lyric into 10 Possible Moral Foundations as defined in Moral Foundation Theory The available Moral Foundations are:* {`Moral Foundations Tags`}. *The explanation of the moral foundations is as follows:* {`Description Tags`}. *This is a multi-label classification problem: where it's possible to assign one or multiple categories simultaneously. Report the results in JSON format such that the keys of the correct moral values are reported in a list.*

The song lyrics utilised for the GPT-4 model classification are the same as the ones annotated by human annotators. In this way, we can compare the human annotations with those of the model while assessing the general performance of GPT-4 for the classification task.

4. EXPERIMENTS

We started by analysing the MoralBERT technique [20] and fine-tuned models using social media data from Twitter, Reddit, and Facebook. The total number of text records was 35,887. We found that 51% of the texts were neutral and 49% of them were labeled with one or more moral values. This indicated a significant skew towards neutral texts, which we addressed by adding the class weighting technique. After that, we evaluated the BERT models fine-tuned with only GPT-4 generated lyrics. We call these models "BERT SL". We also fine-tuned the models with a combination of out-of-domain social media data and the generated lyrics data which we call "MoralBERT SL". We used the Domain Adversarial module only when fine-tuning BERT with multiple domain data, including synthetic lyrics. When fine-tuning solely with synthetic lyrics, this module was not utilized. Lastly, we evaluated GPT-4's zero-shot classification capabilities against our models on the manually annotated song lyrics.

The results show that the models achieving the highest

	F1 Scores Weighted Average				F1 Scores Binary			
	MoralBERT	GPT-4	BERT SL	MoralBERT SL	MoralBERT	GPT-4	BERT SL	MoralBERT SL
Care	.80 ± .03	.68 ± .03	.81 ± .03	**.83 ± .03**	.68 ± .05	.64 ± .04	.68 ± .05	**.75 ± .04**
Harm	.68 ± .03	**.75 ± .03**	.71 ± .03	.70 ± .03	.62 ± .05	**.71 ± .04**	.63 ± .05	.69 ± .04
Fairness	.55 ± .03	.73 ± .03	.73 ± .03	**.74 ± .03**	.30 ± .05	.39 ± .06	**.41 ± .06**	.38 ± .06
Cheating	.84 ± .03	.80 ± .03	**.86 ± .02**	.69 ± .03	.27 ± .09	.16 ± .07	**.52 ± .08**	.32 ± .06
Loyalty	.69 ± .03	.67 ± .03	.77 ± .04	**.79 ± .04**	**.38 ± .06**	.34 ± .06	.21 ± .08	.27 ± .09
Betrayal	.81 ± .02	.72 ± .03	**.89 ± .02**	.84 ± .02	.34 ± .07	.31 ± .06	**.40 ± .11**	.37 ± .08
Authority	.77 ± .03	.75 ± .03	.77 ± .03	**.84 ± .03**	**.45 ± .06**	.42 ± .06	.35 ± .07	.39 ± .09
Subversion	**.80 ± .03**	.72 ± .03	**.80 ± .03**	.71 ± .03	**.44 ± .07**	.39 ± .06	.40 ± .07	.43 ± .06
Purity	.77 ± .03	.86 ± .02	.89 ± .02	**.90 ± .02**	.41 ± .06	.56 ± .07	.55 ± .08	**.63 ± .08**
Degradation	.74 ± .03	.81 ± .03	.81 ± .03	**.86 ± .03**	.34 ± .06	**.40 ± .07**	.30 ± .07	.32 ± .10
Average	.75 ± .03	.75 ± .03	**.80 ± .03**	**.80 ± .03**	.42 ± .06	.43 ± .06	.45 ± .07	**.46 ± .07**

Table 1. F1 scores of prediction models with standard deviation estimated via 1,000 bootstraps. Weighted average scores account for both moral and non-moral (neutral) classes, while binary scores only for moral classes. SL = Synthetic Lyrics.

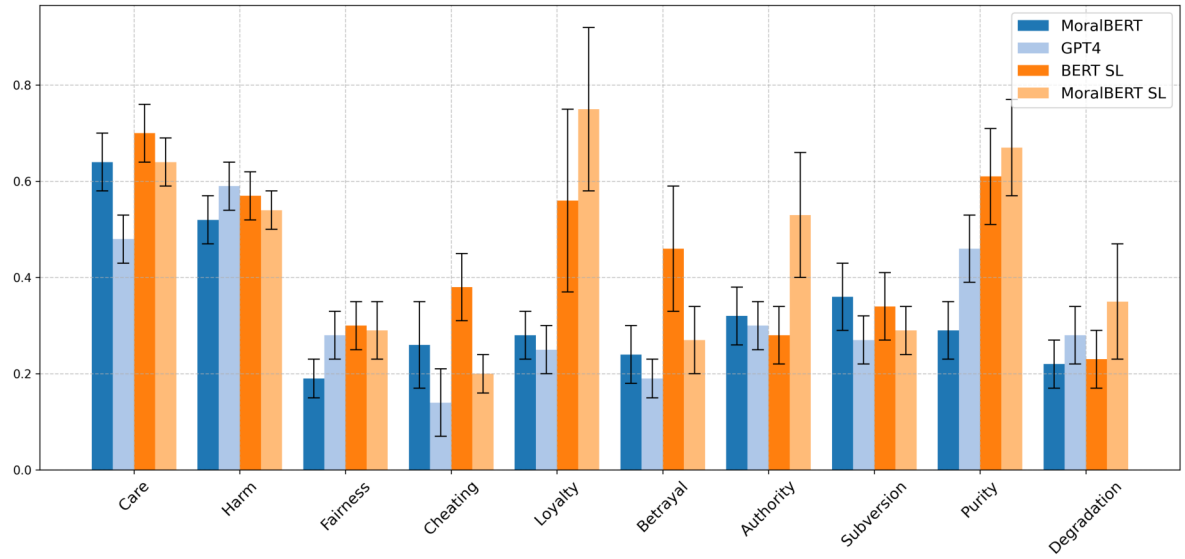

Figure 3. Precision scores for binary classification with standard deviation estimated via 1,000 bootstraps.

performance were BERT SL and MoralBERT SL. These models performed on average 5% better across all moral values in terms of F1 weighted score which accounts for both moral and non-moral prediction classes. While for the binary F1, these models were marginally better than GPT-4. For harm foundation, GPT-4 performed slightly better, possibly due to the synthetic lyrics' lack of natural variability when expressing this foundation. The fact that MoralBERT SL and BERT SL performances are similar to the one from GPT-4 for binary F1 is expected as the same latent knowledge of GPT-4 has been distilled into BERT by using the generated lyrics. The improvements from MoralBERT SL and BERT SL are significant for what concerns weighted F1, suggesting that given the supervised setting of these models, they were also able to learn the higher prior probability of non-moral (e.g., neutral) instances, which generally outweigh moral instances. The same is evident if we look at Figure 3, which compares the binary Precision scores of the various models. From the figure, it is evident that MoralBERT SL and BERT SL exhibit significantly higher Precision surpassing GPT-4 and MoralBERT by 12% on average. These models, then, are often correct when labelling lyrics with moral values (even though results vary according to which moral value), while being more cautious in assigning a moral value, given the preponderance of neutral cases. For the evaluation metrics, we report the standard deviation estimated via Bootstrapping which is a statistical resampling technique used to estimate the variability of the metrics. We used 1,000 bootstraps which is typically sufficient to achieve a reasonable approximation of the standard deviation.

Our findings show that BERT-based models are still comprehensible with larger models such as GPT-4, when fine-tuned properly they can excel in specified tasks. GPT-4 demonstrated a very good performance even without any fine-tuning (zero-shot approach) which was anticipated given its state-of-art performance in multiple tasks and its training on an extensive amount of data. These models have been trained on diverse text sources such as Wikipedia, GitHub, chat logs, books, and articles [41], enabling them to comprehend language across various domains [31]. The earlier model, GPT-3, contains 175 bil-

Song Name	Artist	Human Annotations	MoralBERT	GPT-4	BERT SL MoralBERT SL
"Take This Heart of Mine"	Foghat	Care, Purity	Care, Purity	Care, Loyalty	Care, Fairness, Purity
"Who's Cheatin' Who"	Charly McClain	Cheating, Betrayal	Cheating, Betrayal, Loyalty, Purity	Cheating, Betrayal	Cheating, Betrayal
"Samurai Showdown"	RZA	Harm, Authority	Harm, Betrayal, Authority, Purity	Harm, Loyalty, Authority	Harm, Authority
"Man In The Mirror"	Mark Chesnutt	Care, Fairness	Fairness, Loyalty, Authority	Care, Fairness, Loyalty, Authority	Care, Fairness

Table 2. Examples of moral values detected in song lyrics by human annotators and model predictions.

lion parameters, far exceeding BERT base model with 110 million parameters [42]. Such models demand significantly more computational resources than BERT models. In contrast, the BERT model is cost-free, easier to modify, and offers greater control over the models due to its open-source nature. On the other hand, BERT models need fine-tuning, which presents its own challenges due to the necessity for manual labelling and data annotation. Therefore, a hybrid approach like the one we suggest offers an optimised solution that combines the best of both worlds.

Table 2 presents four song examples annotated for moral values by both human annotators and prediction models. These examples show that MoralBERT SL and BERT SL (not shown in the table as it shares the same outcomes as MoralBERT SL for these instances) aligned most closely with human moral assessments. From a general observation of the song lyrics that were annotated by humans and tested with these models, it was noted that MoralBERT and GPT-4 tend to assign more moral attributes per song while increasing their chances of correctly guessing moral labels but also misclassifying neutral ones. In contrast, models trained with synthetic lyrics more accurately identified neutral (non-moral) lyrics, aligning with the quantitative observations of the F1 weighted score. Typically, human annotators did not assign more than three moral values per song. To control the number of assigned moral values per song, we adjusted the thresholds [37] for our prediction models, ensuring optimal accuracy. When lacking ground truth data, a post-processing can be applied for cutting moral labels with lower probabilities. Here we present only F1 and Precision scores. For further details, refer to the project's results page on GitHub. [2]

5. CONCLUSION

In this paper, we presented an integrated approach for the automatic detection of moral values in lyrics. We created a synthetic lyrics dataset using GPT-4 which we used to fine-tune the BERT-base model alone (BERT SL) and in combination with out-of-domain social media corpora (MoralBERT SL). We introduced a dataset of 200 song lyrics sourced from the WASABI dataset annotated for moral values by two experts, serving as the basis for evaluating our moral prediction models. We also assessed the performance of models trained with synthetic lyrics in comparison to those trained solely on social media data (MoralBERT) and a zero-shot GPT-4 classifier. We found that models trained with synthetic lyrics generally achieved significantly better binary Precision and higher weighted F1 scores compared to the GPT-4 classifier and MoralBERT, along with marginally better binary F1.

Our research has some limitations. To begin with, the synthetic lyrics is created via GPT-4, a powerful model but not an open-source, which limits our control of the model. We prompted GPT-4 to create unique lyrics in the style of various artists across different genres. Yet, adding musical composition details, lyrical themes [43], or visual images as descriptors [44], could enhance both the quality and diversity of the generated lyrics. However, we only employ this method for fine-tuning to make BERT models learn the structure and moral expressions in lyrics. The creation of truly creative lyrics for artistic purposes requires greater sophistication and rigorous human review [44]. Further, we analysed the overall moral expressions in the song lyrics without differentiating between structural elements such as verses, bridges, and choruses. Lastly, we focus on inferring moral values in English lyrics, which limits our ability to understand moral expressions in music lyrics from non-Western cultures.

Understanding how lyrics can convey moral values is important for the MIR field, as it can enhance how we experience and interact with music, including improving music tagging and recommendation systems [45]. Addressing challenges in automatic detection of moral values in lyrics can further push the boundaries of current technologies in natural language processing and machine learning applied to music and other creative tasks. Further, as lyrics often reflect societal values and cultural norms, tools for extracting morality rapidly from lyrical text enable researchers to gain insights into the prevailing moral attitudes of different times or cultures. This can be useful in sociological studies, helping scholars understand how music influences and is influenced by societal norms and changes.

[2] https://github.com/vjosapreniqi/ismir-mft-values/tree/main/Results

6. ETHICS STATEMENT

In this study, we employed large language models (LLMs) to generate synthetic lyrics. Given the vast amount of data on which these models are trained, there is a potential for bias transfer from the training datasets. Additionally, these models may inadvertently contain copyrighted literary works within their training data, necessitating meticulous steps to prevent plagiarism, particularly if the generated lyrics are utilised beyond fine-tuning for artistic and creative outputs [46, 47].

We engaged two human annotators to label 200 songs with moral values based on the Moral Foundations Theory (MFT). These annotators signed a consent document that detailed the project's objectives, their roles, and the nature of their tasks. They were informed of their right to withdraw from the study at any time without consequences. To protect their privacy, all data from the annotators were anonymised.

While powerfull language models like BERT and GPT-4 offer significant potential to enhance communication and support social campaigns, they also pose risks if used for manipulative purposes. Our research is committed to advancing the understanding of moral expressions in music and fostering the responsible development and use of AI in creative contexts.

7. ACKNOWLEDGEMENTS

VP and IG are supported by PhD studentships from Queen Mary University of London's Centre for Doctoral Training in Data-informed Audience-centric Media Engineering. KK acknowledges support from the Lagrange Project of the Institute for Scientific Interchange Foundation (ISI Foundation) which is funded by Fondazione Cassa di Risparmio di Torino (Fondazione CRT).

8. REFERENCES

[1] M. E. Ballard and S. Coates, "The immediate effects of homicidal, suicidal, and nonviolent heavy metal and rap songs on the moods of college students," *Youth & Society*, vol. 27, no. 2, pp. 148–168, 1995.

[2] S. Frith, *Sound effects; youth, leisure, and the politics of rock'n'roll.* Pantheon Books, 1981.

[3] L. Betti, C. Abrate, and A. Kaltenbrunner, "Large scale analysis of gender bias and sexism in song lyrics," *EPJ Data Science*, vol. 12, no. 1, p. 10, 2023.

[4] D. R. Dewberry and J. H. Millen, "Music as rhetoric: Popular music in presidential campaigns," *Atlantic Journal of Communication*, vol. 22, no. 2, pp. 81–92, 2014.

[5] E. J. Kizer, "Protest song lyrics as rhetoric," *Popular Music & Society*, vol. 9, no. 1, pp. 3–11, 1983.

[6] J. O. Adebayo, "Vote not Fight: Examining music's role in fostering non-violent elections in Nigeria," *African Journal on Conflict Resolution*, vol. 17, no. 1, pp. 55–77, 2017.

[7] D. D. Sellnow, "Music as persuasion: Refuting hegemonic masculinity in "He Thinks He'll Keep Her"," *Women's Studies in Communication*, vol. 22, no. 1, pp. 66–84, 1999.

[8] R. Mayer and A. Rauber, "Musical genre classification by ensembles of audio and lyrics features," in *Proceedings of the International Society for Music Information Retrieval (ISMIR)*, 2011, pp. 675–680.

[9] R. Delbouys, R. Hennequin, F. Piccoli, J. Royo-Letelier, and M. Moussallam, "Music mood detection based on audio and lyrics with deep neural net," in *Proceedings of the International Society for Music Information Retrieval (ISMIR)*, 2018, pp. 370–375.

[10] Y. Song and D. Beck, "Modeling emotion dynamics in song lyrics with state space models," *Transactions of the Association for Computational Linguistics*, vol. 11, pp. 157–175, 2023.

[11] A. Vaglio, R. Hennequin, M. Moussallam, G. Richard, and F. d'Alché Buc, "Multilingual lyrics-to-audio alignment," in *Proceedings of the International Society for Music Information Retrieval (ISMIR)*, 2020, pp. 512–519.

[12] N. L. Masclef, A. Vaglio, and M. Moussallam, "User-centered evaluation of lyrics-to-audio alignment." in *Proceedings of the International Society for Music Information Retrieval (ISMIR)*, 2021, pp. 420–427.

[13] J. Kim, A. M. Demetriou, S. Manolios, M. S. Tavella, and C. C. Liem, "Butter lyrics over hominy grit: Comparing audio and psychology-based text features in mir tasks." in *Proceedings of the International Society for Music Information Retrieval (ISMIR)*, 2020, pp. 861–868.

[14] V. Preniqi, K. Kalimeri, and C. Saitis, "Soundscapes of morality: Linking music preferences and moral values through lyrics and audio," *PLOS One*, 2023.

[15] V. Preniqi, I. Ghinassi, C. Ive, Juliaand Saitis, and K. Kalimeri, "Moralbert: A fine-tuned language model for capturing moral values in social discussions," in *ACM 4th International Conference on Information Technology for Social Good (GoodIT)*, 2024.

[16] S. Guo, N. Mokhberian, and K. Lerman, "A data fusion framework for multi-domain morality learning," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 17, 2023, pp. 281–291.

[17] V. Preniqi, K. Kalimeri, and C. Saitis, ""More Than Words": Linking music preferences and moral values through lyrics," in *Proceedings of the International Society for Music Information Retrieval (ISMIR)*, 2022, pp. 797–805.

[18] J. Haidt and J. Graham, "When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize," *Social Justice Research*, vol. 20, no. 1, pp. 98–116, 2007.

[19] J. Hoover, K. Johnson, R. Boghrati, J. Graham, and M. Dehghani, "Moral framing and charitable donation: Integrating exploratory social media analyses and confirmatory experimentation," *Collabra: Psychology*, vol. 4, no. 1, 2018.

[20] V. Preniqi, K. Kalimeri, and C. Saitis, "Modelling moral traits with music listening preferences and demographics," *Music in the AI Era. CMMR 2021. Lecture Notes in Computer Science*, vol. 13770, pp. 183—-194, 2021.

[21] K. J. Messick and B. E. Aranda, "The role of moral reasoning & personality in explaining lyrical preferences," *PLOS One*, vol. 15, no. 1, p. e0228057, 2020.

[22] K. Kalimeri, M. G. Beiró, M. Delfino, R. Raleigh, and C. Cattuto, "Predicting demographics, moral foundations, and human values from digital behaviours," *Computers in Human Behavior*, vol. 92, pp. 428–445, 2019.

[23] O. Araque, L. Gatti, and K. Kalimeri, "MoralStrength: Exploiting a moral lexicon and embedding similarity for moral foundations prediction," *Knowledge-Based Systems*, vol. 191, pp. 1–11, 2020.

[24] F. R. Hopp, J. T. Fisher, D. Cornell, R. Huskey, and R. Weber, "The extended moral foundations dictionary (eMFD): Development and applications of a crowdsourced approach to extracting moral intuitions from text," *Behavior Research Methods*, vol. 53, pp. 232–246, 2021.

[25] J. Graham, J. Haidt, and B. A. Nosek, "Liberals and conservatives rely on different sets of moral foundations," *Journal of Personality and Social Psychology*, vol. 96, no. 5, pp. 1029–1046, 2009.

[26] J. Hoover, G. Portillo-Wightman, L. Yeh, S. Havaldar, A. M. Davani, Y. Lin, B. Kennedy, M. Atari, Z. Kamel, M. Mendlen *et al.*, "Moral foundations Twitter corpus: A collection of 35k tweets annotated for moral sentiment," *Social Psychological and Personality Science*, vol. 11, no. 8, pp. 1057–1071, 2020.

[27] J. Trager, A. S. Ziabari, A. M. Davani, P. Golazazian, F. Karimi-Malekabadi, A. Omrani, Z. Li, B. Kennedy, N. K. Reimer, M. Reyes *et al.*, "The moral foundations Reddit corpus," *arXiv preprint arXiv:2208.05545*, 2022.

[28] M. G. Beiró, J. D'Ignazi, V. Perez Bustos, M. F. Prado, and K. Kalimeri, "Moral narratives around the vaccination debate on Facebook," in *Proceedings of the ACM Web Conference 2023*, 2023, pp. 4134–4141.

[29] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *Proceedings of International Conference on Machine Learning*, 2015, pp. 1180–1189.

[30] E. Liscio, O. Araque, L. Gatti, I. Constantinescu, C. Jonker, K. Kalimeri, and P. K. Murukannaiah, "What does a text classifier learn about morality? an explainable method for cross-domain comparison of moral rhetoric," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2023, pp. 14 113–14 132.

[31] S. Doh, K. Choi, J. Lee, and J. Nam, "LP-MusicCaps: Llm-based pseudo music captioning," in *Proceedings of the International Society for Music Information Retrieval (ISMIR)*, 2023.

[32] Z. Zhang, K. Lasocki, Y. Yu, and A. Takasu, "Syllable-level lyrics generation from melody exploiting character-level language model," in *Findings of the Association for Computational Linguistics: EACL 2024*, 2024, pp. 1336–1346.

[33] P. Sawicki, M. Grzes, L. F. Góes, D. Brown, M. Peeperkorn, A. Khatun, and S. Paraskevopoulou, "On the power of special-purpose GPT models to create and evaluate new poetry in old styles," *University of Leicester*, 2023.

[34] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, pp. 159–174, 1977.

[35] G. Meseguer-Brocal, G. Peeters, G. Pellerin, M. Buffa, E. Cabrio, C. Faron Zucker, A. Giboin, I. Mirbel, R. Hennequin, M. Moussallam *et al.*, "WASABI: A two million song database project with audio and cultural metadata plus webaudio enhanced client applications," *Web Audio Conference (WAC)*, 2017.

[36] G. King and L. Zeng, "Logistic regression in rare events data," *Political Analysis*, vol. 9, no. 2, pp. 137–163, 2001.

[37] O. Koshorek, A. Cohen, N. Mor, M. Rotman, and J. Berant, "Text segmentation as a supervised learning task," *arXiv preprint arXiv:1803.09337*, 2018.

[38] J. Ye, J. Gao, Q. Li, H. Xu, J. Feng, Z. Wu, T. Yu, and L. Kong, "ZeroGen: Efficient zero-shot learning via dataset generation," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Y. Goldberg, Z. Kozareva, and Y. Zhang, Eds. Association for Computational Linguistics, Dec. 2022, pp. 11 653–11 669.

[39] X. He, I. Nassar, J. Kiros, G. Haffari, and M. Norouzi, "Generate, annotate, and learn: Nlp with synthetic text," *Transactions of the Association for Computational Linguistics*, vol. 10, pp. 826–842, 2022.

[40] M. O. Silva, L. M. Rocha, and M. M. Moro, "MusicOSet: An enhanced open dataset for music data mining," in *XXXII Simpósio Brasileiro de Banco de Dados: Dataset Showcase Workshop, SBBD*, 2019, pp. 8–17, Accessed online: 2024-02-05.

[41] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877–1901, 2020.

[42] M. Bosley, M. Jacobs-Harukawa, H. Licht, and A. Hoyle, "Do we still need BERT in the age of GPT? comparing the benefits of domain-adaptation and in-context-learning approaches to using llms for political science research," *University of Michigan*, 2023.

[43] K. Watanabe, Y. Matsubayashi, K. Inui, T. Nakano, S. Fukayama, and M. Goto, "Lyrisys: An interactive support system for writing lyrics based on topic transition," in *Proceedings of the 22nd International Conference on Intelligent User Interfaces*, 2017, pp. 559–563.

[44] K. Watanabe and M. Goto, "Text-to-lyrics generation with image-based semantics and reduced risk of plagiarism," in *Proceedings of the International Society for Music Information Retrieval (ISMIR)*, 2023.

[45] A. Laplante, "Improving music recommender systems: What can we learn from research on music tastes?" in *in Proceedings of the International Society for Music Information Retrieval (ISMIR)*, 2014, pp. 451–456.

[46] J. Barnett, "The ethical implications of generative audio models: A systematic literature review," in *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 2023, pp. 146–161.

[47] F. Morreale, M. Sharma, I. Wei *et al.*, "Data collection in music generation training sets: A critical analysis," in *Proceedings of the International Society for Music Information Retrieval (ISMIR)*, 2023.