Positive and Risky Message Assessment for Music Products

Yigeng Zhang[1], Mahsa Shafaei[1], Fabio A. González[2], Thamar Solorio[1,3]
[1]University of Houston, Houston, USA
[2]Universidad Nacional de Colombia, Bogotá, Colombia
[3]MBZUAI, Masdar City, United Arab Emirates
[1]{yzhang168,mshafaei,tsolorio}@uh.edu
[2]fagonzalezo@unal.edu.co

Abstract

In this work, we introduce a pioneering research challenge: evaluating positive and potentially harmful messages within music products. We initiate by setting a multi-faceted, multi-task benchmark for music content assessment. Subsequently, we introduce an efficient multi-task predictive model fortified with ordinality-enforcement to address this challenge. Our findings reveal that the proposed method not only significantly outperforms robust task-specific alternatives but also possesses the capability to assess multiple aspects simultaneously. Furthermore, through detailed case studies, where we employed Large Language Models (LLMs) as surrogates for content assessment, we provide valuable insights to inform and guide future research on this topic. The code for dataset creation and model implementation is publicly available at https://github.com/RiTUAL-UH/music-message-assessment.

Keywords: Document Classification, Text Categorization, Text Mining, Tools, Systems, Applications

1. Introduction

Accessing music has never been more convenient than it is today. People can use various tools, such as high-fidelity players and streaming apps, to enjoy music at any time. Listeners can simply go online, press the *PLAY* button, and find themselves invigorated after a bad day. However, this easy access also raises concerns that children and adolescents may have a higher chance of being exposed to risky content. Young people's thoughts and behavior might potentially be affected by the positive or questionable content in songs, as they tend to learn from the modeled behavior that popular music represents (Primack et al., 2008). For example, a study has shown an association between adolescent early sexual experiences and degrading sexual content in music (Primack et al., 2009). Similarly, researchers have also revealed that listening to particular types of songs is positively associated with substance use and aggressive behaviors (Chen et al., 2006).

The American Academy of Pediatrics (AAP) holds the opinion that parents should be informed of pediatricians' concerns regarding the potentially harmful effects of music lyrics (American Academy of Pediatrics, 1996). Policymakers have also taken action; for instance, the Recording Industry Association of America (RIAA) introduced the Parental Advisory Label (PAL) program in 1985 to identify audio products with potentially inappropriate content. Such labels were created to draw parents' attention to products that may not be suitable for their children. Recent studies from the NLP community have made advances in automating the content rating process. Chin et al. and Fell et al. presented machine learning-based methods to automatically classify explicit/non-explicit lyrics in different languages. Further work (Rospocher and Eksir, 2023) studied detecting explicitness in several aspects such as strong language and language that refers to violence. The learning objectives are derived from content rating systems like PAL, and these works treat the problem as a binary classification task. Despite these existing achievements, however, neither the current PAL system nor the corresponding automated approaches are capable of appraising a music product with severity information about content suitability.

In this work, we introduce a novel NLP task: assessing the positive and risky messages of a music item. We study the messages that a music item conveys from five significant dimensions regarding appropriateness: *Positive Messages*, *Violence*, *Substance Consumption*, *Sex*, and *Consumerism* along three degrees of severity. In light of those perspectives, we propose a multi-task method that successfully incorporates both content aspect correlations and severity ordinalities to better assess music products. Our research focuses specifically on music performed with vocal techniques (singing, rapping, etc). In such a format of music, the lyrics play a dominant part in conveying the message and opinion from the artists, which has a major impact on the listener's thoughts and minds. Leveraging an automated approach for music content assessment can offer numerous advantages to various stakeholders in the music industry. For music providers, it promises enhanced service quality, particularly benefiting younger listeners and their

LREC-COLING 2024, pages 12893–12905

guardians by facilitating better content recommendations and advisories. Furthermore, lyricists can utilize this system during the early stages of their creative process to assess the portrayal of potentially risky behaviors, such as *'get drunk and be somebody'*. Additionally, the automated identification of questionable content can expedite the creation of clean lyric versions, enabling artists to produce safe renditions more efficiently.

Our contribution: To the best of our knowledge, this study represents the first exploration into assessing music items, evaluating both their positive and risky dimensions across multiple levels. We have established a comprehensive benchmark and have made the dataset creation method accessible to the wider research community[1]. Additionally, we introduce an effective multi-task, ordinality-enforced rating approach that jointly assesses diverse risk factors, delivering state-of-the-art results. As part of our thorough analysis, we also evaluate the efficacy of Large Language Models (LLMs) and provide in-depth discussions, paving the way for future research for this topic.

2. Music rating and lyrics

To build up a reliable benchmark for assessing positive and risky messages in such lyrics, we exploit expert ratings provided by Common Sense Media (CSM)[2]. CSM is a non-profit organization caring for kids' digital well-being. It hosts a media rating platform that is supported by childhood development experts. The rating system covers various aspects of childhood development and age appropriateness. In this work, based on the focal points from Youth Risk Behavior Surveillance System (YRBSS) (Kolbe et al., 1993) introduced by the USA Centers for Disease Control and Prevention, we choose three explicitly risky aspects *Violence*, *Substance Consumption*, and *Sex*. We also pick *Consumerism* that relates to drawing interest in the acquisition of goods, often associated with a "feel good" experience. Because *Consumerism* potentially affects the emotional health and identity development in youth (Hill, 2011). Our dataset also includes a category dedicated to *Positive Messages* within music products. The automatic detection of such uplifting content offers numerous valuable applications, enhancing user experience. This could serve as a counter to the often conflicting messages found in mainstream lyrics. Furthermore, it might be leveraged as a form of supportive therapy to positively influence users' moods.

[1] https://github.com/RiTUAL-UH/music-message-assessment.

[2] https://www.commonsensemedia.org.

2.1. Dataset facts and analysis

With CSM expert ratings, we developed a dataset based on publicly available lyrics from the Internet. We collected 1,119 music items (consisting of 10,661 songs) including standard albums, extended plays (EP), long plays (LP), and CD singles. We refer to the collections of songs as *albums* in this work for simplification. The partition of CD singles and the albums are shown in Table 1. The *age recommendation* spans from 2 to 18+. Notably, the music items we consider in this work are in English and comprise a heavily Western-centric dataset.

Category	Album	CD Single
Percentage	62.2%	37.8%

Table 1: Percentage of albums and CD singles.

The duration of individual music items, including both CD singles and albums, shows significant variation. Comprehensive distributions of text lengths are illustrated in Figure 1.

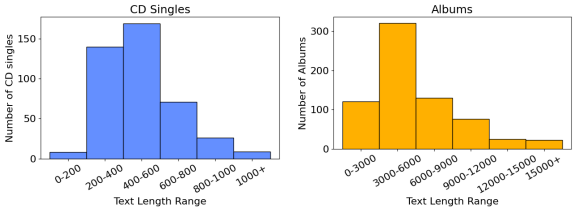

Figure 1: Lyrics length distribution for CD singles and albums.

The level of aspect prevalence is from 0 to 5. To reduce class imbalance due to the fact of data availability, we then project such scored ratings into 3-level ordinal categories (Low Presence (0-1), Medium Presence (2-3), High Presence (4-5)) with median split strategy as in related works (Martinez et al., 2019, 2020) from the movie domain. A detailed label distribution is described in Table 2.

The music rating data comes from Common Sense Media (CSM). Non-member users can only browse up to three expert product reviews for free each month. We gained permission from CSM to use the music rating data for research, however, by the time we submit this work, the music reviews are no longer shown on CSM's website until further updates. We still release the names of expert-rated products and links to their lyrics studied in this work for the community. We do not release the lyrics directly due to copyright considerations; however, lyrics can be easily accessed through the links we provide and via search engines. While we've made every effort to collect lyrics from the open Internet, we acknowledge that some songs might be missing from certain albums.

Aspect	Violence	Substance	Sex	Consumerism	Positive
Low	844	743	663	880	736
Medium	190	294	319	209	314
High	85	82	137	30	69

Table 2: Low/medium/high presence item distribution for each message aspect.

2.1.1. Inter-dimensional correlation analysis

Empirically, risky behaviors such as violence and substance use often appear concurrently in a piece of music. We further calculate the Spearman rank correlation ρ between each positive and risky rating pair among all of the music items. The correlation heat map is demonstrated in Figure 2. All correlation scores between variable pairs have significance with $p < .05$ except *Positive Messages-Consumerism*.

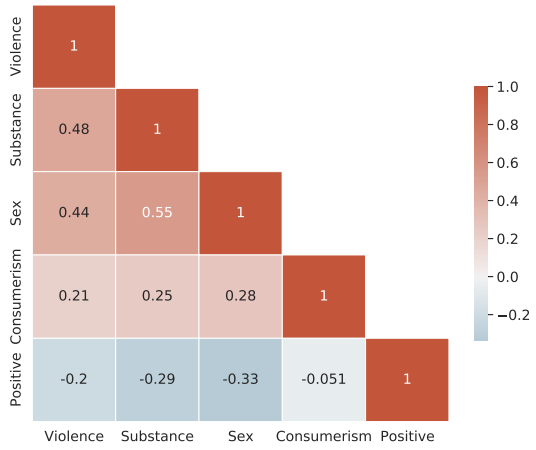

Figure 2: Spearman rank correlation between positive and risky prevalence pairs.

We can easily observe from the heat map that typical physical risky behaviors *Violence*, *Substance Consumption*, and *Sex* have a positive correlation with each other, while *Consumerism* has a positive correlation with those three but less strong. It is no surprise that *Positive Messages* has a significant negative correlation with three physical risky behaviors. This interdimensional behavior intuitively inspires us to leverage such correlations to design relevant machine learning strategies.

2.1.2. Positive/Risky behavior *v.s.* Explicitness

In the context of music, *explicitness* often refers to content that contains strong language or violence, sex, or substance abuse depictions (from RIAA). It is a generalized description that has a high coincidence with the risky message we studied in this work. Since it is not practical to collect gold labels of explicitness for every music item we studied, we apply an explicit lyrics classifier trained on 438k English lyrics from a previous work (Fell et al., 2020). We can further explore the correlation between the explicitness probability and the level of positive and risky messages. Figure 3 illustrates the pattern of explicitness probability towards severity levels of different aspects.

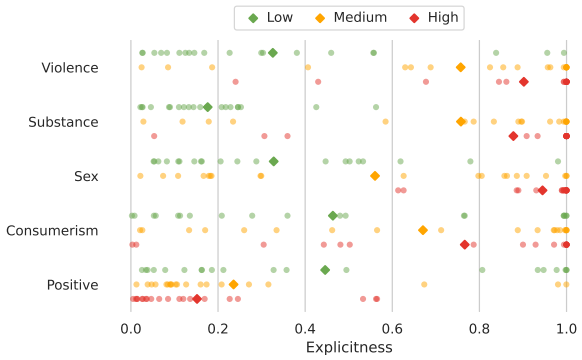

Figure 3: Explicitness probability on different message dimensions. 20 instances are randomly sampled from each level. The diamond symbol ◆ indicates the central tendency of a series of probability values for the corresponding level.

The message ratings consistently correlate with explicitness values. For all risky message aspects, explicitness increases as ratings transition from low to high. Even the abstract message of *Consumerism* aligns with this trend. In contrast, *Positive Messages* exhibit a distinct negative correlation with explicitness. These correlations suggest that message ratings can serve as a valuable complement or alternative to traditional metrics of music appropriateness. They provide a richer perspective and highlight the significance of our study.

3. Methodology

We formulate this music content assessment problem as a multi-class text classification task, where the ratings of an aspect are the prediction objectives. We take solely lyrics as visible features. Expecting to leverage the correlation pattern among different messages and their ordinal levels, we propose an effective model that incorporates rich semantic representation, aspect-aware multi-task learning, and ordinality-enforcement. The model architecture is illustrated in Figure 4.

Emotion-guided twin model: We propose to encode the text from two perspectives: general-

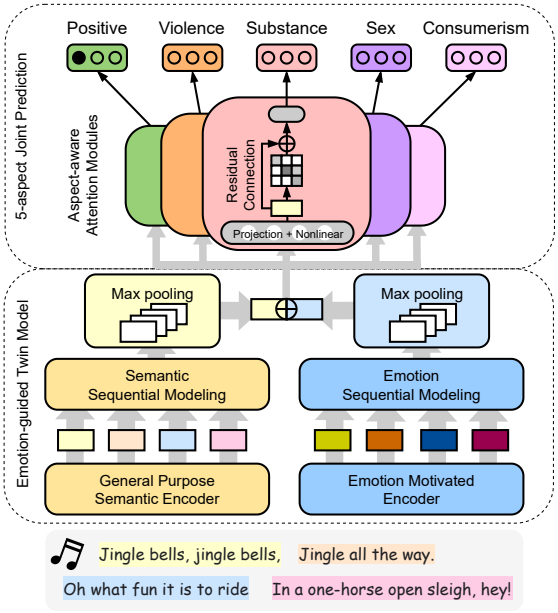

Figure 4: Joint prediction architecture with emotion-guided twin and aspect-aware attention module.

purpose semantic representation and emotion-centered semantic representation, as emotion information has been proved to be an effective complement for various language understanding tasks (Shafaei et al., 2020; Samghabadi et al., 2020). We employ two pretrained Transformer models fine-tuned with semantic textual similarity (Sentence-BERT (Reimers and Gurevych, 2019)) and emotion detection (task-specific Distilled RoBERTa on emotion detection (Hartmann, 2022)) respectively. The final representation is the concatenation of these two types of embeddings.

Joint prediction over multiple aspects: From the dataset study, we found different types of positive and risky messages, specifically questionable behaviors, have a potential correlation with each other. For instance, the presence of violence in a music item might appear concurrently with lyrics depicting substance use. The learning objective is a joint loss of predicting ratings of multiple aspects. To fortify the representation uniqueness for different aspects, we further design an aspect-aware attention module to learn specialized features for each aspect. The module begins with a projection layer with non-linearity and further learns weights through an aspect-differentiation matrix. The final representation $\boldsymbol{x}_{out}$ comes from the addition of input $\boldsymbol{x}_{in}$ with residual over skip connections as described in Equation 1.

$$\boldsymbol{x}_{out} = \boldsymbol{x}_{in} + \boldsymbol{x}_{in} \circ Softmax\left(\boldsymbol{x}_{in}\boldsymbol{W}_{attn}\right) \quad (1)$$

Ordinality-enforcement techniques: The severity ratings are discrete interclass ordinal variables instead of independent categorical labels. Typical classification models usually ignore such correlations between ordinal categories. In this work, we apply three ordinality-enforcement techniques that are better suited for the task. All of the ordinality-enforcement techniques are applied respectively to the base model with a multi-tasking module.

- ***Siamese ranking-classification*** (Zhang et al., 2021): This method leverages a Siamese network to process a pair of instances for both ranking and classification objectives. The ranking (comparison) operation has the potential to learn pairwise ordinal differences in severity levels between samples. The model is optimized with two cross-entropy losses for the two objectives as in Equation 2, where l_{cls} comes from multi-class classification while l_{rank} is derived from comparing ratings (lower/same/higher) between the two music items.

$$\hat{f} \leftarrow \arg\min_{f} \left(l_{\text{cls}} + l_{\text{rank}}\right) \quad (2)$$

- ***Binary attributes transformation*** (Frank and Hall, 2001): This method tackles ordinal regression by dividing the sorted ordinal label set, containing n elements, into two subsets at every possible pair of adjacent elements. This results in $n-1$ potential splits. Each split transforms the ordinal regression problem into a binary classification task, where the goal is to predict whether the ordinal value y_i falls before or after the split point within the set. This approach applies multiple classifiers to leverage the ordinal information repeatedly. Specifically, in the setting of this problem with three ordinal classes (low, medium, and high), we consider two binary splits: one between low and medium, and another between medium and high. This results in two binary classifications:

$$\begin{aligned} &\Pr(y_i \leqslant \text{Low}), \Pr(y_i > \text{Low}); \\ &\Pr(y_i \leqslant \text{Mid}), \Pr(y_i > \text{Mid}). \end{aligned} \quad (3)$$

  In this setting, for each classification, we apply a binary classifier to predict which split the prediction will fall in, utilizing the ordinal information multiple times aiming to improve the prediction performance.

- ***Soft label*** (Diaz and Marathe, 2019): This method introduces a label softening function to convert ordinal category values into a probability distribution across categories. A class label y_i from the label set is encoded into a soft label c_i^{soft} using the following formula for a specific true rating rank y_t.

$$c_i^{soft} = \frac{e^{-\phi(y_t, y_i)}}{\sum_{k=1}^{K} e^{-\phi(y_t, y_k)}} \quad \forall y_i \in \mathbb{Y}. \quad (4)$$

The Kullback-Leibler divergence is used as the loss function to measure the difference between the predicted probability and the soft label. $\phi(y_t, y_i) = |y_t - y_i|$ is chosen as the metric penalty for the sake of simplicity.

4. Experiments and Results

To evaluate the effectiveness of our proposed method in music content assessment, we benchmarked several popularly used classification methods and models from related works in media rating. We choose TF-IDF and Bag-of-Word-Vectors (Averaged GloVe (Pennington et al., 2014) embeddings) with SVM classifiers, TextCNN (Kim, 2014) and TextRCNN (Lai et al., 2015). We experiment with three deep models that are designed for media rating problems:

1. An RNN model with attention for predicting movie MPAA ratings based on movie dialogue scripts (Shafaei et al., 2020) as lyrics are also sequences of utterances;

2. A BERT model (Devlin et al., 2019) as used for classifying explicitness in (Fell et al., 2020);

3. An RNN-Transformer backbone model of state-of-the-art in rating severity for age-restricted content in movies (Zhang et al., 2021).

The training-development-test split uses an 80/10/10 ratio with data shuffled. We choose macro F1 as the classification performance metric because the label distribution from the dataset is highly imbalanced. Experimental results are shown in Table 3.

Among deep baseline models, there is no dominant architecture that can give the best prediction performance on every aspect but the RNN-Transformer (RT) model has an overall best performance with a notable gap. Our proposed method, emotion-guided multi-tasking model with ordinality-enforcement, shows an overall best performance among all methods. The repeated-measures t-test shows our ordinality-enforcement models give a significant performance improvement ($p < .05$) over the strongest baseline on average, specifically on explicit risky behaviors (*Violence*, *Substance consumption*, and *Sex*), using five random seeds with 10-fold cross-validation mean F1 score.

5. Discussion and analysis

Speaking of aspects, three types of explicit content, *Violence*, *Substance*, and *Sex* are relatively easier for the model to predict than *Consumerism* and *Positive messages*. The reason could be that the latter aspects are more abstract concepts compared to risky behaviors that can be described in the lyrics in an overt manner. For *Consumerism*, we hypothesize that a model which can more effectively capture signals related to 'goods' and the promotion of purchases might perform better. Unlike explicit questionable content like *Violence*, *Positive messages* is even harder to intuitively find a clear textual pattern since *positive* is more heterogeneous than the other aspects. We may need the model to gain a high-level understanding of the content to distinguish the quantity of positive value that a song or an album delivers.

5.1. Ablation study

We conducted an ablation study on the best-performing model by iteratively removing individual components from the model architecture. The experiment demonstrated that all modules within the network structure contributed to the performance in prediction. When we configured the proposed network for single-task prediction with all modules, the performance reached even higher levels. This indicates that the effectiveness of the novel network modules can help single-task baseline models learn better quality text representations. Additionally, we experimented with the backbone model by having it predict all five aspects directly in a multi-task setting, which resulted in a large performance drop. This finding highlights the effectiveness and necessity of incorporating components that can guide and enhance multitask learning.

5.2. Saliency analysis

We perform saliency analysis using input perturbation to better understand the model prediction behavior. We chose a segment from the lyrics of *Heartless (2019)* by *The Weekend*, which contains explicit language. We do the perturbation sentence by sentence through removing one and feeding the rest of the lyrics into the model, then we inspect the model prediction result. Table 5 shows the detailed influence on prediction confidence and outcome. For *Substance Consumption*, line 1 explicitly contains the word *drunk*. When we remove this sentence, the prediction drastically changes from *High* to *Low*. For *Sex*, when we remove the first line, the prediction probability decreases and the result becomes *Low*. The same situation happens to the fourth line. Interestingly, removing line 2 results in an upgrade. We hypothesize that this deletion increases the density of sexual implication content. This case study intuitively shows the model can successfully capture particular mentions of risky messages such as substance use and sex-related topics that have significance in severity prediction.

Baselines	Violence	Substance	Sex	Consumerism	Positive	Avg
Majority voting	28.65	26.59	24.76	30.22	26.46	27.34
TF-IDF + SVM	51.28	47.31	61.16	40.51	27.79	45.61
BoWV + SVM	46.75	40.23	55.63	31.15	28.28	40.41
TextCNN (Kim, 2014)	50.59	48.25	59.03	44.88	36.79	47.91
TextRCNN (Lai et al., 2015)	57.63	55.86	64.01	44.61	38.47	52.12
LSTM+Attention (Shafaei et al., 2020)	33.93	34.48	41.65	34.62	30.06	34.95
BERT (Devlin et al., 2019; Fell et al., 2020)	56.93	52.16	60.61	44.96	**46.57**	52.25
RNN Trans (RT) (Zhang et al., 2021)	62.84	62.47	67.36	46.16	44.37	56.64
multi-task + ordinality-enforcement						
Soft label	61.42	64.36	68.92	45.21	42.60	56.50
Ranking-classification	**65.04**	**64.41**	**69.11**	45.65	44.72	57.79
Binary transformation	64.46	63.98	69.00	**47.36**	44.61	**57.88**

Table 3: Experimental result on positive and risky message level with macro F1 scores with 10-fold cross-validation.

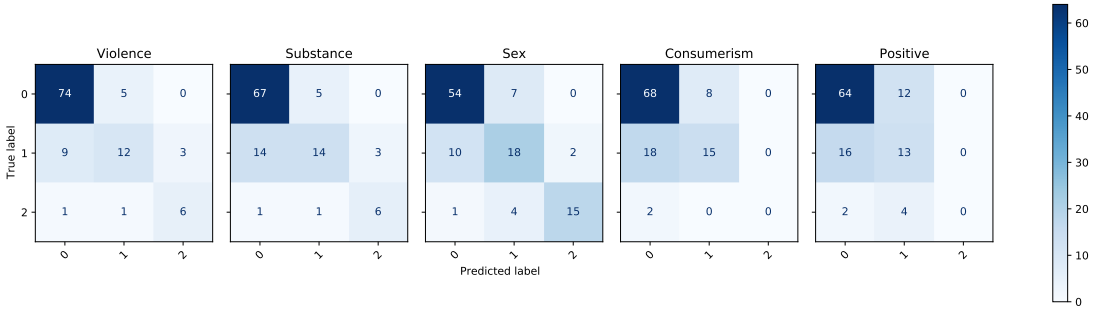

Figure 5: Prediction confusion matrix of the best-performing method on all 5 aspects. The x-axis indicates the predicted values and the y-axis indicates the ground-truth labels.

Ablation	Perf change
Best performing model	57.88
Aspect-aware module	57.58 (-0.30)
Emotion-guided twin model	57.77 (-0.11)
Ordinality enforcement (binary)	57.53 (-0.24)
Multitask joint prediction	59.05 (1.17)
Backbone only + multitask	55.36 (-2.52)

Table 4: Ablation study of different components in the best-performing model. We report and analyze the average performance changes across the five aspects.

5.3. Error analysis

Figure 5 shows the confusion matrix of the prediction results of the best-performing model from a single fold of the cross-validation. In general, the proposed method can capture the ordinal information well because wrong predictions that cross two levels (predict low to high or high to low) are rare. Specifically, the model struggled to give correct predictions on high *Consumerism*. We suppose the number of training instances of high *Consumerism* is relatively small. The same case happened to high *Positive* predictions. We hypothesize the heterogeneous nature of *Positive* content makes it challenging to predict.

5.4. Case study: unsuccessful predictions

We dig into some unsuccessful predictions to analyze the errors. We mainly focus on the hard aspects of the model.

- **Album: The Best Damn Thing (2007) by Avril Lavigne**: The proposed model gives *Consumerism* low and *Positive* low ratings, however, the correct labels for those two aspects are medium. This album is a pop-punk production and the songs in the album seemed to be targeted at young people with themes such as love and encouragement. For *Consumerism*, there are explicit lyrics saying:

  > *I hate it when a guy doesn't get the tab*
  >
  > *And I have to pull my money out, and that looks bad*

  But such cases are rare and one will not make a confident decision for a strict medium rating. For *Positive*, this album contains lyrics with significant positive values such as *Keep Holding On*:

  > *You're not alone*
  >
  > *Together we stand*

	Removed sentence from a segment of *Heartless (2019)*	Violence	Substance	Sex
	Rating	Low	High	Mid
	Whole segment confidence	98.77	90.57	72.60
1	Stix drunk, but he never miss a target	1.13	-90.56 ⇓	-64.57 ↓
2	Photoshoots, I'm a star now (Star)	-4.18	2.48	-40.27 ↑
3	I'm talkin' Time, Rolling Stone, and Bazaar now (Bazaar now)	-1.91	1.55	-15.39
4	Sellin' dreams to these girls with their guard down (What?)	-1.71	2.01	-30.90 ↓

Table 5: An input perturbation study on the behavior of the proposed ranking-classification model. We choose three risky aspects - *Violence*, *Substance consumption*, and *Sex* - as this model yields the best performance. The numbers indicate the absolute probability change of the original prediction result. A double down arrow ⇓ indicates the predicted severity downgraded by two levels, a single down arrow ↓ means downgraded by one, and an up arrow ↑ represents upgraded by one.

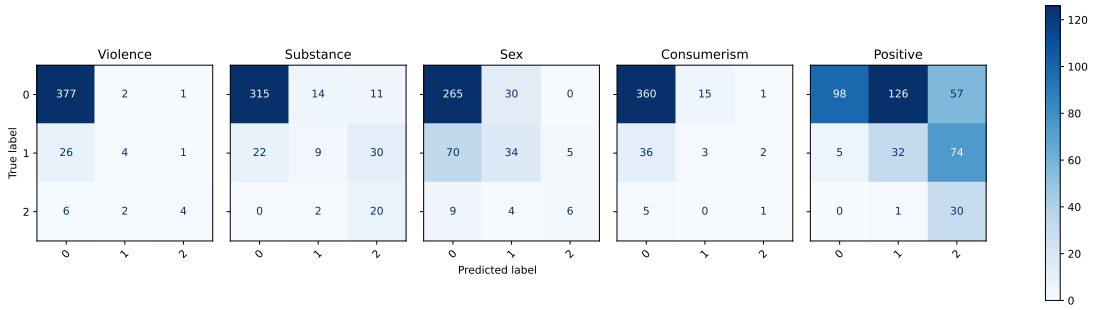

Figure 6: Prediction confusion matrix of the best-performing LLM method on all 5 aspects. The x-axis indicates the predicted values and the y-axis indicates the ground-truth labels.

> *I'll be by your side, you know I'll take your hand*

We suspect that explicit expressions of positivity as in the example are sparse in the songs. However, many songs in the album convey a spirit of pursuing happiness and love. This nuance gives the album a medium rating in *Positive*, but it is challenging for the model to capture.

- **Album: A Hard Day's Night (1964) by The Beatles**: The model gives a critical wrong prediction on *Positive*: predict as low while the ground truth is high. The majority of the songs in this album express deep and genuine emotions about love, with lyrics like:

> *If you need somebody to love, just look into my eyes*
> *I'll be there to make you feel right*
> *If you're feeling sorry and sad, I'd really sympathize*
> *Don't you be sad, just call me tonight*

One possible reason the model failed is that the lyrics do not contain words that explicitly convey strong positive signals of companionship. However, the overall sentiment is clearly positive and constructive.

- **CD single: Labels or Love (2008) by Fergie**: The model gives *Sex* medium, *Substance consumption* medium, and *Consumerism* low, while the correct answer are *Sex* low, *Substance consumption* low, and *Consumerism* high. For *Sex* medium and *Substance consumption*, the false positive may come from lexical signals such as *sexy*, *kiss* and *bag*. For *Consumerism*, there are not only expressions to the encouragement of buying goods:

> *Let's stop chasing those boys and shop some more*

but also many explicit mentions of luxury brands:

> *Gucci, Fendi, Prada purses, purchasing them finer things*
> *Men, they come a dime a dozen*
> *Just give me them diamond rings*
> *I'm into a lot of bling, Cadillac, Chanel and Coach*

We suspect that the model did not interpret a direct narrative about shopping as a strong indicator of *Consumerism*. Additionally, the names of luxury brands might be rare in the corpus, leading to a lack of supervision signals. As a result, the model struggled with this prediction.

	Vio	Sub	Sex	Con	Pos	Avg
Simple context	54.91	49.87	49.09	45.35	34.35	46.71
Rich context	53.51	**54.06**	53.79	40.94	**35.33**	**47.53**
Simple context + CoT	53.72	37.83	43.22	**45.95**	29.77	42.10
Rich context + CoT	**55.52**	50.18	**54.03**	40.82	29.05	45.92

Table 6: Zero-shot evaluation results from `gpt-3.5-turbo`. Only CD-singles are evaluated due to the intrinsic token size limitations of this LLM.

5.5. Case study: LLM as content judges

Recent advancements in Large Language Models (LLMs) have showcased impressive natural language understanding and adaptability across a multitude of tasks. Motivated by these advancements, our study aims to explore the potential of LLMs in assessing content within music products. Specifically, we leverage the `gpt-3.5-turbo` API (OpenAI, 2022; Ouyang et al., 2022) as surrogate evaluators. Our primary focus is on rating five specific aspects of the content. The central hypothesis of this study is that, despite their inherent limitations and lack of access to a supervision signal, LLMs can provide content assessments that are both meaningful and comparable in accuracy to other deep learning methods. Due to the context length constraints of LLMs, we limit our evaluation to CD-singles, setting a token cap of 3000 to ensure the model's efficient functioning.

The experiments were structured in three distinct formats:

- Simple Context: The LLM is directed to rate each song across the five aspects without any supplementary information.
- Rich Context: Before prompting the LLM, a detailed description of the five aspects is provided in the context.
- Chain-of-Thought (CoT): Building upon the CoT approach (Wei et al., 2022), known for enhancing LLM performance in complex reasoning tasks, we feed the model with exemplar prompts as context and subsequently instruct it to complete the rating task.

Table 6 presents the evaluation results. While LLM-based approaches have their merits, they did not demonstrate remarkable efficiency in this content rating context. Although providing a richer context yielded marginally superior outcomes, no method consistently outperformed the others. Interestingly, LLM assessments aligned with patterns observed in our baseline and proposed methods. Specifically, the LLM found it more straightforward to evaluate the explicit aspects of *Violence*, *Substance*, and *Sex*, but faced challenges with *Consumerism* and *Positive*. It's important to note that this experimental design is simpler than previous sections, given it exclusively assesses CD-singles rather than an assorted selection from an album.

Further analysis of the LLM's performance was conducted by examining the confusion matrix of the top-performing model, as depicted in Figure 6. The model displayed some notable patterns in its behavior. Specifically:

- For categories like *Violence*, *Sex*, and *Consumerism*, the model tended to underestimate their respective severities.
- When rating *Substance*, the model frequently struggled to give *medium presence* ratings.
- In contrast, the evaluation of *Positive* content often resulted in an overestimation of a song's positive messages, different from patterns observed in prior experiments of baseline and proposed methods.

It's essential to recognize the intrinsic limitations of the LLM. We could not apply the same training and assessment methods to the LLM as we did in previous sections. Although CD singles are a subset of the broader music product collection, they retain a consistent data property. While our comparison does not strictly align with traditional comparative analysis standards, due to potential disparities in data distribution and features across datasets, it nonetheless provides valuable insights. These insights can guide model benchmarking and optimization, even if not strictly empirical. Our decision to evaluate closed models like `gpt-3.5-turbo`—which is non-reproducible due to its proprietary nature—stems from a desire to explore the capabilities of such models. We urge readers to interpret these particular results as exploratory, rather than as fixed benchmarks.

While LLMs have demonstrated proficiency in a variety of NLP tasks, their performance in our specialized context of content assessment was not on par. This discrepancy is understandable given that these versatile models are not trained for such tasks. Consequently, their judgments might not always align with the expert opinions of professionals in media research and childhood development. Recognizing this, our forthcoming research aims to explore the development and analysis of task-specific LLMs for content safety. We are optimistic

that such an approach will yield meaningful insights and enhanced performance.

6. Conclusion

In this paper, we introduce a novel task to the NLP community: predicting the intensity of various aspects of music, spanning from objectionable content to positive messages. By analyzing music product lyrics, we investigate multiple dimensions of messages conveyed to listeners. Our research problem and approach are intended to foster deeper investigations into music content assessment. The multi-task ordinality-enforcement model we present has shown promising effectiveness for this type of challenge with ordinal properties. The case studies, along with our exploration using Large Language Models (LLMs) as surrogate evaluators, highlight the inherent complexities of the message assessment problem, calling for the need for continued community engagement and research.

Ethical considerations and limitations

This work introduces a novel task: assessing positive and risky messages for music products. It also proposes a state-of-the-art method to automatically accomplish the assessment. We acknowledge the potential limitations and ethical considerations by highlighting the following points for future explorations on similar topics:

Reliability: We recognize the potential issues regarding the reliability of such a content rating system. Possible inaccuracies may result in misleading content suggestions, potentially leading vulnerable groups to inadvertently consume inappropriate content, or causing confusion in the production processes for musicians. This work represents our initial exploration, and we strongly advise against implementing such a system in real-world services until the technical and operational elements can be held accountable. We insist that such a system should be regarded as an assistant to, rather than a replacement for, the content rating and assessment work done by media experts and customers.

Social context concerns: The social context in which these labels are acquired is not always known, and there can be a lack of context in terms of how language is used and judged. For instance, many rap songs discuss the harms associated with isolation and substance abuse, yet such information might be misclassified due to rating provider bias or system bias. This may increase the likelihood of systematic bias or unintentionally promote racism. Future research should explore distinguishing between racist and reclaimed uses of slurs as well as between mentions of risky subjects in a suggestive manner and those deemed more innocuous.

Data source concerns: Our rating data comes from CSM, a non-neutral organization. The regulations they use to recruit human experts to rate media products are unknown, and different experts rate different products, which may lead to inconsistencies. Transparency and accountability are not guaranteed, and subjectivity remains in the rating results. Future research and implementations should not rashly take ratings from data sources such as CSM as golden standards without careful assessment.

Ambiguity in aspects: The rating aspects used in this work can be ambiguous, as they are loosely defined by single words. For example, what is considered *violence* may extend beyond overtly violent behaviors. We also recognize that the *positive* aspect defined by the CSM, indicating the overall takeaway, lacks fine-grained elaborations compared to risky aspects. This could result in ambiguity and fail to provide further insights for users.

Implications and acceptable use: An ethical concern of this study is the potential for media censorship. We acknowledge that efficient machine learning-based algorithms like the proposed method could be used as censorship tools. Malicious users could misuse the proposed method for illegitimate censorship, potentially harming freedom of speech. We call for the development and use of AI algorithms with special attention to who should use the system, how it should be used, and what safeguards should be in place to prevent misuse.

Modality coverage: A technical limitation of this work is that it does not take other modalities of music products, such as melody, rhythm, and vocal performance, into account for predictions. We recognize the significance of these signals in conveying a song's message. Additionally, we focused only on songs in English in this study, which means lyrics written in other languages and from different cultural backgrounds are absent from our study. Our work stands as an initial exploration of the feasibility of solving this task. We hope that future work can explore diversifying the dataset and exploring how the models behave in those cases.

Acknowledgements

We thank Common Sense Media for permitting us to use the expert ratings for research. We would like to thank the anonymous LREC-COLING reviewers for their feedback on this work.

7. Bibliographical References

Alfred V. Aho and Jeffrey D. Ullman. 1972. *The Theory of Parsing, Translation and Compiling*, volume 1. Prentice-Hall, Englewood Cliffs, NJ.

American Academy of Pediatrics. 1996. Impact of music lyrics and music videos on children and youth. *Pediatrics*, 98(6):1219–1221.

American Psychological Association. 1983. *Publications Manual*. American Psychological Association, Washington, DC.

Rie Kubota Ando and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853.

Galen Andrew and Jianfeng Gao. 2007. Scalable training of L1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, pages 33–40.

BSI. 1973a. *Natural Fibre Twines*, 3rd edition. British Standards Institution, London. BS 2570.

BSI. 1973b. Natural fibre twines. BS 2570, British Standards Institution, London. 3rd. edn.

A. Castor and L. E. Pollux. 1992. The use of user modelling to guide inference and learning. *Applied Intelligence*, 2(1):37–53.

Ashok K. Chandra, Dexter C. Kozen, and Larry J. Stockmeyer. 1981. Alternation. *Journal of the Association for Computing Machinery*, 28(1):114–133.

Meng-Jinn Chen, Brenda A Miller, Joel W Grube, and Elizabeth D Waiters. 2006. Music, substance use, and aggression. *Journal of studies on alcohol*, 67(3):373–381.

J.L. Chercheur. 1994. *Case-Based Reasoning*, 2nd edition. Morgan Kaufman Publishers, San Mateo, CA.

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

N. Chomsky. 1973. Conditions on transformations. In *A festschrift for Morris Halle*, New York. Holt, Rinehart & Winston.

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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Raul Diaz and Amit Marathe. 2019. Soft labels for ordinal regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4738–4747.

Umberto Eco. 1990. *The Limits of Interpretation*. Indian University Press.

Michael Fell, Elena Cabrio, Elmahdi Korfed, Michel Buffa, and Fabien Gandon. 2020. Love me, love me, say (and write!) that you love me: Enriching the wasabi song corpus with lyrics annotations. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2138–2147.

Eibe Frank and Mark Hall. 2001. A simple approach to ordinal classification. In *European conference on machine learning*, pages 145–156. Springer.

Dan Gusfield. 1997. *Algorithms on Strings, Trees and Sequences*. Cambridge University Press, Cambridge, UK.

Jochen Hartmann. 2022. Emotion english distilroberta-base. https://huggingface.co/j-hartmann/emotion-english-distilroberta-base/.

Jennifer Ann Hill. 2011. Endangered childhoods: How consumerism is impacting child and youth identity. *Media, Culture & Society*, 33(3):347–362.

Paul Gerhard Hoel. 1971a. *Elementary Statistics*, 3rd edition. Wiley series in probability and mathematical statistics. Wiley, New York, Chichester. ISBN 0 471 40300.

Paul Gerhard Hoel. 1971b. *Elementary Statistics*, 3rd edition, Wiley series in probability and mathematical statistics, pages 19–33. Wiley, New York, Chichester. ISBN 0 471 40300.

Otto Jespersen. 1922. *Language: Its Nature, Development, and Origin*. Allen and Unwin.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

Lloyd J Kolbe, Laura Kann, and Janet L Collins. 1993. Overview of the youth risk behavior surveillance system. *Public Health Reports*, 108(Suppl 1):2.

Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao. 2015. Recurrent convolutional neural networks for text classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(1).

Victor Martinez, Krishna Somandepalli, Yalda Tehranian-Uhls, and Shrikanth Narayanan. 2020. Joint estimation and analysis of risk behavior ratings in movie scripts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4780–4790, Online. Association for Computational Linguistics.

Victor R Martinez, Krishna Somandepalli, Karan Singla, Anil Ramakrishna, Yalda T Uhls, and Shrikanth Narayanan. 2019. Violence rating prediction from movie scripts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 671–678.

OpenAI. 2022. Introducing chatgpt. Accessed: 2023-10-14.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Brian A Primack, Madeline A Dalton, Mary V Carroll, Aaron A Agarwal, and Michael J Fine. 2008. Content analysis of tobacco, alcohol, and other drugs in popular music. *Archives of pediatrics & adolescent medicine*, 162(2):169–175.

Brian A Primack, Erika L Douglas, Michael J Fine, and Madeline A Dalton. 2009. Exposure to sexual lyrics and sexual experience among urban adolescents. *American journal of preventive medicine*, 36(4):317–323.

Mohammad Sadegh Rasooli and Joel R. Tetreault. 2015. Yara parser: A fast and accurate dependency parser. *Computing Research Repository*, arXiv:1503.06733. Version 2.

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)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Marco Rospocher and Samaneh Eksir. 2023. Assessing fine-grained explicitness of song lyrics. *Information*, 14(3).

Sebastian Ruder. 2017. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*.

Niloofar Safi Samghabadi, Afsheen Hatami, Mahsa Shafaei, Sudipta Kar, and Thamar Solorio. 2020. Attending the emotions to detect online abusive language. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 79–88.

Mahsa Shafaei, Niloofar Safi Samghabadi, Sudipta Kar, and Thamar Solorio. 2020. Age suitability rating: Predicting the MPAA rating based on movie dialogues. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1327–1335, Marseille, France. European Language Resources Association.

Charles Joseph Singer, E. J. Holmyard, and A. R. Hall, editors. 1954–58. *A history of technology*. Oxford University Press, London. 5 vol.

Jannik Strötgen and Michael Gertz. 2012. Temporal tagging on different domains: Challenges, strategies, and gold standards. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 3746–3753, Istanbul, Turkey. European Language Resource Association (ELRA).

S. Superman, B. Batman, C. Catwoman, and S. Spiderman. 2000. *Superheroes experiences with books*, 20th edition. The Phantom Editors Associates, Gotham City.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Yigeng Zhang, Mahsa Shafaei, Fabio Gonzalez, and Thamar Solorio. 2021. From none to severe: Predicting severity in movie scripts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3951–3956, Punta Cana,

Dominican Republic. Association for Computational Linguistics.

A. Implementation details

We list the implementation details of the proposed baseline methods.

Sparse and dense document representation: Here we apply TF-IDF and Bag-of-Word-Vectors (GloVe (Pennington et al., 2014) Average) as text vectorization approaches and linear models as baselines. TF-IDF method is also used in the previous work that classifies explicitness in lyrics (Fell et al., 2020).

Word-level semantic representation: Word vectors such as Word2Vec and Glove are effective semantic representations of NLP tasks. We apply TextCNN (Kim, 2014) with Glove embedding as a benchmark. The Glove embedding vectors used in the experiments are trained on Wikipedia 2014 and Gigaword 5, with 300 dimensions. The TextCNN has kernel sizes of 3, 4, and 5 in the convolution modules.

Word-level semantic representation with sequence modeling: With word vectors, we further utilize the TextRCNN (Lai et al., 2015) model to capture the sequential signal out of each individual word. TextRCNN and other RNN-based models utilize a bi-directional LSTM structure with hidden sizes of 200.

Word-level sequence modeling with attention mechanism: This model performs the best in predicting the MPAA ratings based on movie scripts and rich metadata (Shafaei et al., 2020). For script text processing, they apply LSTM to model the sequential information from the word embeddings and use attention mechanism to aggregate the output of each time step for text representation.

Pretrained language model task fine-tuning: Contextualized representation from pretrained Transformer-based models have shown significant success on various NLP tasks. One popular variant, BERT (Devlin et al., 2019), was also used in classifying explicitness in the previous work(Fell et al., 2020). We fine-tune a pretrained BERT on this task in the multi-class classification setting. The BERT model is adapted from HuggingFace with a maximum input length limitation of 512 tokens.

Sentence-level semantic representation with sequence modeling: This model is the state-of-the-art in a severity rating problem for age-restricted content in movies (Zhang et al., 2021). We leverage the strong representation capability from the pretrained languages to obtain semantic representations. Then we apply general-purpose sentence embedding from Sentence-BERT (Reimers and Gurevych, 2019) to encode each sentence from the lyrics of a music item. Then the semantic representation sequences are further encoded using a recurrent architecture to model sequential information. It also becomes a part of the backbone model in our proposed method.

Emotion-guided Transformer model: We apply a Distilled RoBERTa model that is finetuned on emotion detection tasks (Hartmann, 2022) to obtain the emotion-guided sentence embeddings.

Multitask and ordinality-enforcement: Our proposed multitask model predicts 5 aspects (*Positive Messages*, *Violence*, *Substance Consumption* (Drinking, drugs, and smoking), *Sex*, and *Consumerism*) at one single prediction. The ordinality-enforcement components are applied to each individual aspect prediction.

All experiments are conducted using NVIDIA Tesla P40 and PyTorch 1.6.0/PyTorch Lightning 1.0.2. The optimizer is Adam optimizer with 0.001 as the learning rate. Each training epoch of the proposed method takes less than 30 seconds under a batch size of 40.

A.1. Ranking-classification loss behavior

The training behavior of the ranking-classification joint loss in the ordinality-enforcement method is shown in Figure 7. Both cross-entropy losses are averaged on each loss instance in one batch. The ranking loss is often higher during the training process. We suppose ranking is more challenging because the ranking pairs are randomly constructed for every new training step.

A.2. Prompt used for LLM evaluation

Simple context: `Please assess the lyrics of the song given the following aspects: The lyrics: <Full Lyrics> Positive Messages; Violence & Scariness; Sex, Romance & Nudity; Drinking, Drugs & Smoking; Products & Purchases (refer to Consumerism). Please rate the presence of each aspect on a scale of 0 to 2, where 0 indicates 'low', 1 indicates 'medium', and 2 indicates 'high'. Provide your ratings strictly in the JSON format as shown in the example below and make sure no extra content: Example: {"Positive Messages": 1, "Violence & Scariness": 0, "Sex, Romance & Nudity": 1, "Drinking, Drugs & Smoking": 2, "Products & Purchases": 1}.`

Rich context: `Please assess the lyrics of the song given the following aspects: The lyrics: <Full Lyrics> Positive Messages: <Full aspect`

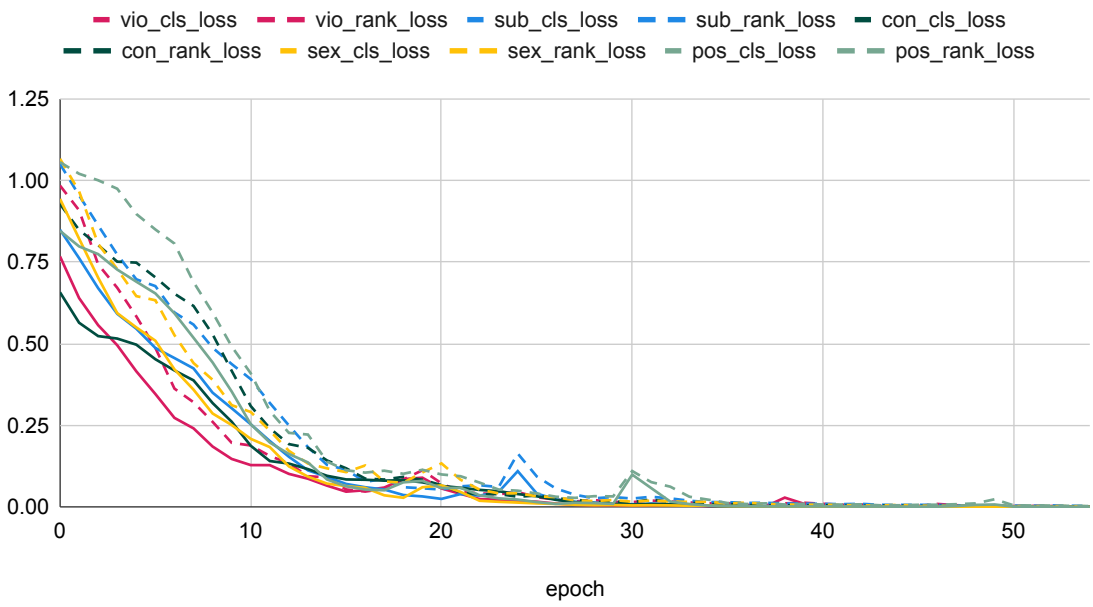

Figure 7: Training loss behavior of the ranking-classification method. The classification and ranking losses within one aspect are grouped by the same tone of the color palette: classification losses are in solid lines while ranking losses are in dash lines.

description from CSM>; Violence & Scariness: <Full aspect description from CSM>; Sex, Romance & Nudity: <Full aspect description from CSM>; Drinking, Drugs & Smoking: <Full aspect description from CSM>; Products & Purchases (refer to Consumerism): <Full aspect description from CSM>. Please rate the presence of each aspect on a scale of 0 to 2, where 0 indicates 'low', 1 indicates 'medium', and 2 indicates 'high'. Provide your ratings strictly in the JSON format as shown in the example below and make sure no extra content: Example: {"Positive Messages": 1, "Violence & Scariness": 0, "Sex, Romance & Nudity": 1, "Drinking, Drugs & Smoking": 2, "Products & Purchases": 1}.

Chain-of-Thought (CoT) prompt:

(Keep context part the same). Please rate the presence of each aspect on a scale of 0 to 2, where 0 indicates 'low', 1 indicates 'medium', and 2 indicates 'high'. Let's think step-by-step to analyze the lyrics and then provide your ratings in the JSON format. Here is an example output: This song promotes ... and the song has strong ... It depicts ..., so it implies ... (your chain-of-thought) ... Therefore, we reach the final assessment result: {"Positive Messages": 1, "Violence & Scariness": 0, "Sex, Romance & Nudity": 1, "Drinking, Drugs & Smoking": 2, "Products & Purchases": 1}. Now it is your turn: