information

MDPI

Article

Assessing Fine-Grained Explicitness of Song Lyrics

Marco Rospocher * and Samaneh Eksir

Department of Foreign Languages and Literatures, University of Verona—Lungadige Porta Vittoria, 41-37129 Verona, Italy

* Correspondence: marco.rospocher@univr.it

Abstract: Music plays a crucial role in our lives, with growing consumption and engagement through streaming services and social media platforms. However, caution is needed for children, who may be exposed to explicit content through songs. Initiatives such as the Parental Advisory Label (PAL) and similar labelling from streaming content providers aim to protect children from harmful content. However, so far, the labelling has been limited to tagging the song as explicit (if so), without providing any additional information on the reasons for the explicitness (e.g., strong language, sexual reference). This paper addresses this issue by developing a system capable of detecting explicit song lyrics and assessing the kind of explicit content detected. The novel contributions of the work include (i) a new dataset of 4000 song lyrics annotated with five possible reasons for content explicitness and (ii) experiments with machine learning classifiers to predict explicitness and the reasons for it. The results demonstrated the feasibility of automatically detecting explicit content and the reasons for explicitness in song lyrics. This work is the first to address explicitness at this level of detail and provides a valuable contribution to the music industry, helping to protect children from exposure to inappropriate content.

Keywords: text classification; multi-label tagging; explicit content detection; natural language processing

check for updates

Citation: Rospocher, M.; Eksir, S. Assessing Fine-Grained Explicitness of Song Lyrics. *Information* **2023**, *14*, 159. https://doi.org/10.3390/info14030159

Academic Editor: Kostas Stefanidis

Received: 8 February 2023
Revised: 28 February 2023
Accepted: 28 February 2023
Published: 2 March 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Music plays a pivotal role in our lives. Recent reports (e.g., IFIP's Engaging with Music 2022 report—https://www.ifpi.org/wp-content/uploads/2022/11/Engaging-with-Music-2022_full-report-1.pdf, accessed on 2 February 2023) show that music consumption keeps growing over the years and people engage with music in many ways, thanks also to the broad diffusion through streaming services and social media platforms. Despite all the benefits music can bring, some caution is necessary when the listeners are children. Songs, as any other form of art, such as movies and photographs, may contain information and messages that may not be adequate for young children or that require at least some proper adult guidance: in these cases, songs are typically referred to as "explicit", meaning that they may use strong or discriminatory language, refer to violence or sex, and so on.

Over the years, several initiatives and organizations, such as the Recording Industry Association of America (RIAA) (https://www.riaa.com/, accessed on 2 February 2023) or the British Phonographic Industry (BPI) (https://www.bpi.co.uk/, accessed on 2 February 2023) have contributed to preventing the exposure of children to content hurtful to them. For instance, RIAA introduced the Parental Advisory Label (PAL), to alert adults that the content of a given music album may be inappropriate for children. Similar label-based strategies are also adopted by streaming content providers (e.g., iTunes, Amazon Music, Spotify, Deezer) to mark songs on their platform that may be explicit, typically with a tag such as "explicit" or "E". However, as this labelling is typically provided as the result of a human (and time-consuming) activity (these tags are typically provided to the content providers by right-holders, record labels, or even end users, cf. https://support.deezer.com/hc/en-gb/articles/360000590898-Explicit-content, accessed on 2

February 2023), systems for automatically tagging explicit songs would greatly benefit the music industry.

In this work, we focused in particular on the explicitness of song lyrics, i.e., the words of a song. In the last few years, different approaches have been presented and compared to automatically tag songs whose textual content could be hurtful or inappropriate for children [1–7]. Given a song's lyrics, all these approaches simply return the information on whether the song is explicit or not. However, in the case that a song is tagged as explicit, they lack any explanation about the reasons for the explicitness, it being, for instance, due to strong language, and/or reference to sex, and/or substance abuse, etc. Having this more precise, fine-grained information on the explicitness of song lyrics could help streaming platforms provide more adequate content to their young users and support parents in making informed decisions on the content their children consume. For instance, different age limits could be set based on the reason for the explicitness, or, based on culture and society, the explicitness of a certain type (e.g., strong language) could be considered inappropriate for children of a certain age in some parts of the world, but tolerable in others.

In this paper, we specifically addressed this problem: Can we develop a system capable of detecting explicit song lyrics and, in the case of explicitness, assessing the reason(s) for the explicitness? To tackle this problem, we framed it as a classification task, where each song's lyrics may be annotated with zero, one, or more labels, each associated with a possible explanation for the explicitness. First, we developed a new dataset of 4000 song lyrics, each manually annotated according to five, well-recognized, possible reasons for explicitness: (i) strong language; (ii) substance abuse; (iii) sexual reference; (iv) reference to violence, physical, or mental abuse; and (v) discriminatory language. Then, we trained and evaluated some machine-learning (ML)-based systems, to assess the feasibility of the tasks. More in detail, the novel contributions of the work are manifold:

1. We propose and released a new dataset of 4000 song lyrics manually annotated with explicitness information. Besides the indication of whether the song lyrics contain explicit content or not, each explicit song's lyrics was also appropriately annotated according to the five reasons for explicitness previously mentioned. To the best of our knowledge, this is: (i) the first released dataset containing manually curated information on the explicitness of song lyrics (the few available datasets mainly rely on explicitness information provided by online platforms (e.g., Spotify in [5]), acknowledged to be inaccurate by the platform themselves (e.g., https://support.spotify.com/us/article/explicit-content/), accessed on 2 February 2023); and (ii) the first dataset containing fine-grained explicitness annotations, detailing the reasons for the explicitness of the song lyrics. The development of datasets is fundamental for the advancement of the state-of-the-art in computer science and related disciplines, especially for problems for which training and testing material is lacking, as the one considered in this paper.
2. We present a preliminary assessment of the quality of the explicitness information available on a popular online streaming platform (Spotify), comparing, on the same songs, the explicitness tags in the platform with our manual annotations.
3. We experimented with some ML classifiers to assess the feasibility of automatically predicting the explicitness and, if so, the reasons for the explicitness of a given song's lyrics. To the best of our knowledge, no previous work has addressed the problem of providing possible reasons for the explicitness of song lyrics. We also released, as part of the evaluation material, a pre-trained model for predicting the explicitness, and possible reasons for it, of any English song lyrics.

The dataset and evaluation materials are made publicly available at https://github.com/rospocher/explicit-lyrics-detection, accessed on 2 February 2023.

The paper is structured as follows. Section 2 presents relevant related work on detecting the explicitness of song lyrics. Section 3 defines the addressed problem and the considered reasons for explicitness. Section 4 illustrates the creation of the novel manually annotated dataset with detailed explicitness information, its main characteristics, and the

relation with existing explicitness information in a well-known online streaming platform (Spotify). Section 5 describes the development and assessment of two systems for automatically predicting the reason(s) for the explicitness of any song lyrics. Finally, Section 7 concludes with some final remarks and hints at directions for future work.

2. Related Work

The focus of this study was on identifying explicit content in song lyrics, a flavour of the general task of detecting offensive content in written text. Therefore, we primarily reviewed and compared previous studies that utilized Natural Language Processing (NLP) methods to identify offensive content in written lyrics. For completeness, we recall that a recent study [8] has also considered the problem of detecting explicit content directly in the audio track of the song, employing an audio-to-character recognition model and a Random Forest classifier to infer explicitness.

Chin et al. [1] studied the problem of lyrics written in the Korean language. A dataset of approximately 28,000 lyrics, with 3.7% being labelled as offensive, was compiled, and various methods were compared on it, from simple techniques such as identifying profanity in a language dictionary to more advanced methods utilizing machine learning classifiers such as AdaBoost, the latter achieving the highest scores.

The identification of explicit lyrics in Korean songs was also addressed by [2], using a dataset of around 70,000 lyrics, where 10.7% were labelled as offensive. Various methods were evaluated and compared such as a lexicon-based filtering approach, powered by a dictionary of explicit words, and a neural model based on a variant of recurrent neural networks (HAN) that uses hierarchical information to process words. The latter method, when combined with vector representations built from the dictionary, achieved the best performance.

In the study by [3], the problem of detecting offensive lyrics in English songs was examined using a dataset of approximately 25,000 lyrics, where 13% were labelled as offensive. The study experimented with multiple machine learning classifiers, including linear Support Vector Machine (SVM) and Random Forest, and preprocessed the lyrics with vectorization techniques such as Term Frequency–Inverse Document Frequency (TF–IDF) and Doc2Vec to extract features for the algorithms. The highest scores were achieved by using Random Forest and TF–IDF vectors.

Fell et al. [9] compared various machine learning and neural classifiers on a dataset of approximately 179,000 English song lyrics, where 9.9% were labelled as offensive. The tested classifiers included Logistic Regression (LR) using TF–IDF vector representations, the BERT Language Model, and Textual Deconvolution Saliency, a Convolutional Neural Network (CNN) for text classification. The study found that deep models performed similarly to simpler approaches, with logistic regression even performing slightly better than BERT. In a subsequent study, the authors created WASABI, an LR classifier using TF–IDF vector representations, trained on around 438,000 English lyrics.

Rospocher [5] proposed to use FASTTEXT word embeddings and its classifier to detect explicit lyrics and evaluated its performance on the largest dataset considered so far, which included approximately 808,000 English lyrics, 7.74% of which were labelled as offensive. The study showed that the FASTTEXT method outperformed several baselines, such as majority voting, logistic regression, and WASABI. Later on, Reference [6] showed that: (i) classifiers built on top of popular transformer-based language models (e.g., BERT [10], RoBERTa [11]) effectively perform for explicit lyrics detection, all achieving state-of-the-art scores, including the smaller and faster DistilBERT [12] approach, making it a competitive pick from the lot; and (ii) classifiers built on top of transformer-based language models do not substantially outperform lighter and computationally less-demanding baselines, such as 1D CNN. A similar comparison, on approximately 34K Italian song lyrics, was reported in [7].

Some observations on these related works can be drawn. First, most of the works (e.g., [2,3,5,7,9]) relied on the explicitness information made available through existing

platforms (e.g., LyricsFind, Deezer, Spotify) which, as previously observed, may not be accurate nor complete. Moreover, the goal of all these works was to determine if a song contains explicit content or not, and none of them attempted to provide a more fine-grained prediction of explicit song lyrics, to identify the possible reason(s) for the explicitness of the lyrics.

3. Problem

The classical problem of detecting explicit song lyrics can be formulated as follows: given the lyrics of a song, determine if they contain content that is inappropriate for or harmful to children. That is, a binary output (explicit/non-explicit) is foreseen.

In this work, we propose to go beyond this mere binary formulation by requiring that, in the case of explicitness, additional details on the reason(s) for the explicitness should be provided. More in detail, we propose to codify the possible distinct reasons for the explicitness of song lyrics so that, given the lyrics of a song, if they are predicted as explicit, one or more of these reasons also needs to be suggested. Inspired also by available materials and guidelines (e.g., https://imusician.pro/en/support/faqs/article/what-is-considered-as-explicit-content, https://soundplate.com/what-does-explicit-content-mean-on-spotify-apple-music-other-music-streaming-platforms/, accessed on 2 February 2023), we considered and adopted the following classification for the possible reasons for the explicitness of song lyrics:

Strong language: The song lyrics include offensive words or curse words, i.e., words generally found to be disturbing and that are not normally used in regular conversation. Swear words (e.g., fuck yourself, bitch) are generally considered strong language. An example of song lyrics containing strong language are those of "Spaz" by "N.E.R.D." (e.g., "I'm a star bitch, I don't give a fuck") (https://www.musixmatch.com/lyrics/N-E-R-D/Spaz, accessed on 2 February 2023).

Substance abuse: The song lyrics refer to excessive use (e.g., getting stoned, getting high, indulging in a dependency) of a drug, alcohol, prescription medicine, etc., in a way that is detrimental to self, society, or both. Both psychological and physical addiction to some substances are covered by this concept. An example of song lyrics referring to substance abuse are those of "Alcohol" by "The Kinks" (e.g., "Who thought I would fall, A slave to demon alcohol") (https://www.musixmatch.com/lyrics/The-Kinks/Alcohol, accessed on 2 February 2023).

Sexual reference: The song lyrics contain references to sex, sexual organs, sexual body parts, sexual activity, sexual abuse, and so on. Idiomatic phrases such as "go fuck yourself" or "what the fuck" were excluded and categorized as strong language instead. An example of song lyrics containing sexual reference are those in "Morning Wood" by "Rodney Carrington" (e.g., "'Cause underneath the covers is my morning wood") (https://www.musixmatch.com/lyrics/Rodney-Carrington/Morning-Wood, accessed on 2 February 2023).

Reference to violence: The song lyrics contain references to hurting a person or living being intentionally, including the description or suggestion of acts typically considered as violent (e.g., killing, stabbing, mentally or physically torturing, committing suicide). Both physical and mental violence and abuse are covered by this concept. Idiomatic expressions using words that are associated with violent acts (e.g., "my heart is bleeding", "I'm devastated") are typically not considered evidence of a reference to violence. An example of song lyrics containing a reference to violence are those of "Story of a Hero" by "Drearylands" (e.g., "You have killed or left dying in the waste") (https://www.musixmatch.com/lyrics/Drearylands/Story-of-a-Hero, accessed on 2 February 2023).

Discriminatory language: The song lyrics contain (i) insulting or pejorative expressions referring to races, ethnic groups, nationalities, genders, sexual orientation, etc.; (ii) offensive language directed at one specific subset of people; (iii) reiteration of stereotypes that can be hurtful for a specific target group of people. An example of song lyrics using discriminatory language are those of "Dash's Interlude" by "Rapper Big Pooh" (e.g., "Faggots gonna

hate me") (https://www.musixmatch.com/lyrics/Rapper-Big-Pooh-feat-O-Dash/Dash-s-Interlude, accessed on 2 February 2023).

Please note that these categories are non-exclusive reasons for the explicitness of a song: the very same song lyrics can be explicit for more than one reason and, thus, be tagged with more than one of these categories.

Given the definition of the problem, we propose to tackle it as a multi-label classification task, a form of classification problem where multiple non-exclusive labels may be assigned to each object to be classified. In the given context, the objects to be classified are the song lyrics, while the multiple non-exclusive labels are the five categories describing the reasons for explicitness: if no label (i.e., reason for explicitness) is assigned to a song's lyrics, the song is considered as non-explicit, while the contrary holds otherwise. By framing the problem this way, we can take advantage of the recent developments in machine-learning-based text classification to address the fine-grained detection of explicit song lyrics.

4. Dataset

To evaluate the performance of any system capable of predicting the reason for the explicitness of a song's lyrics according to the proposed problem formulation, and to possibly support the training of machine learning approaches, a dataset of song lyrics annotated according to the categories presented in Section 3 is needed. No such dataset is available: previously considered approaches have worked with song lyrics just annotated with a binary value concerning the explicitness of the content, relying for the latter on the information, possibly inaccurate and incomplete, provided by available platforms such as LyricsFind, Deezer, or Spotify.

We thus contribute a new dataset of English song lyrics, manually annotated with detailed information on the reasons for the explicitness (if any) of their content. We first collected 4000 song lyrics by randomly sampling one of the datasets previously used to tackle the detection of explicit song lyrics [5], which consisted of approximately 808K song lyrics from LyricWiki automatically annotated with information, provided by Spotify, regarding the presence of explicit content. More precisely, we randomly selected from the latter 2000 song lyrics marked explicit and 2000 song lyrics marked non-explicit. On average, the selected song lyrics were 963 characters long (min: 269; max: 1500).

The song lyrics were collected in a spreadsheet, one per row, enriched with five additional columns, one for each of the aforementioned categories, to be used by the human annotators to mark the corresponding reason(s) for its explicitness (if any). Two annotators were involved in the tagging process, both with C1-level English language proficiency. The annotators were provided with the definition of the categories presented in Section 3 as guidelines for performing the tagging and were asked, for each song's lyrics they processed, to mark (inserting a value of 1) the column(s) that better represent the reasons for its explicitness (if any). If no columns were marked (i.e., all values were 0), the song was considered non-explicit by the annotator.

Initially, 100 song lyrics were tagged by both annotators independently. This initial set was used to compare the tagging quality of both annotators and to assess the feasibility of the annotation process, by computing the Inter-Annotator Agreement (IAA). Both Fleiss's Kappa and Krippendorff's Alpha were calculated, using MASI as the distance metric (cf., [13]), obtaining scores of 0.829 and 0.802 respectively (similar scores were obtained also using the Jaccard metric instead of MASI), thus confirming the substantial agreement between the annotators and the reliability of the annotations produced. Each of the remaining song lyrics was then assigned to a single annotator for tagging. The human effort for annotating the whole dataset was estimated to be approximately 60 h.

4.1. Dataset's Statistics

The resulting dataset consisted of 4000 song lyrics, 1707 of them annotated as explicit and 2293 annotated as non-explicit by the human annotators. Figure 1 (left) shows the number of song lyrics annotated for each of the considered categories.

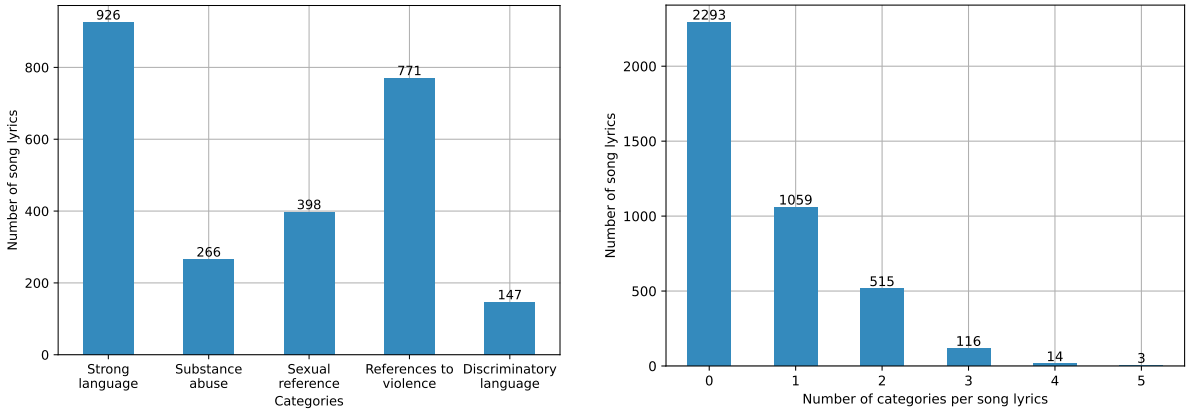

Figure 1. Number of song lyrics annotated for each of the considered categories (**left**); distribution of the song lyrics based on the number of categories they were annotated with (**right**).

By far, the most-popular categories were "Strong Language" and "Reference to Violence", while the other three ones were much less observed.

Figure 1 (right) shows, instead, the distribution of the song lyrics based on the number of categories they were annotated with, ranging from 0 (i.e., the song lyrics are non-explicit) to 5 (i.e., the song lyrics are explicit according to all five proposed categories). It is interesting to observe that more than 25% of the song lyrics were considered explicit for two or more reasons, with a few songs (3) even annotated with all the five considered categories. This demonstrates once again the truly multi-label nature of the problem.

Delving further into this analysis, Table 1 details the co-occurrence matrix between categories, that is the number of times each category (row) was used to annotate the very same lyrics with another category (column). The information on the number of song lyrics annotated with just that single category is also shown (last column, "Single annotation").

Table 1. Co-occurrence counts for reasons for explicitness. Proportions over the total counts of the row category are reported between round brackets. "Single annotation" refers to the counts of the cases where the row category does not co-occur with any other category.

	Strong Language	**Substance Abuse**	**Sexual Reference**	**References to Violence**	**Discriminatory Language**	**Single Annotation**
Strong language	926 (1.00)	109 (0.12)	186 (0.20)	276 (0.30)	99 (0.11)	339 (0.37)
Substance abuse	109 (0.41)	266 (1.00)	47 (0.18)	60 (0.23)	20 (0.08)	103 (0.39)
Sexual reference	186 (0.47)	47 (0.12)	398 (1.00)	89 (0.22)	45 (0.11)	143 (0.36)
References to violence	276 (0.36)	60 (0.08)	89 (0.12)	771 (1.00)	46 (0.06)	403 (0.52)
Discriminatory language	99 (0.67)	20 (0.14)	45 (0.31)	46 (0.31)	147 (1.00)	11 (0.07)

For instance, the table shows that "Reference to violence" occurred very frequently (>50% of the cases) as the only annotation of the song lyrics, while the contrary held for "Discriminatory language", which almost always (93% of the cases) co-occurred with another category, typically "Strong language" (67%); the latter may be due to the fact that, in many cases, swear words or the like are used when offending or insulting a specific target group of people. Indeed, for similar reasons, "Strong language" was the category that most-frequently co-occurred when one of the other categories was used (36–67% of the cases), followed by "Reference to violence" (22–31% of the cases), while the others co-occurred less frequently.

4.2. Comparison with the Explicitness Information Available on Spotify

The song lyrics of the dataset were actually annotated with explicitness information coming from two sources, our manual annotation, as well as the explicitness tags provided by Spotify. We thus deemed it interesting to compare them, especially in light of the note regarding the potential incompleteness of the information about explicitness available on Spotify ("We can't tag all explicit content because it depends on information we get from rights-holders."—https://support.spotify.com/us/article/explicit-content/, accessed on 2 February 2023). Figure 2 (left) shows the distribution of the 4000 song lyrics of the datasets in the four disjoint groups obtained by considering all the combinations of the explicit/non-explicit information available from Spotify and our manual annotation.

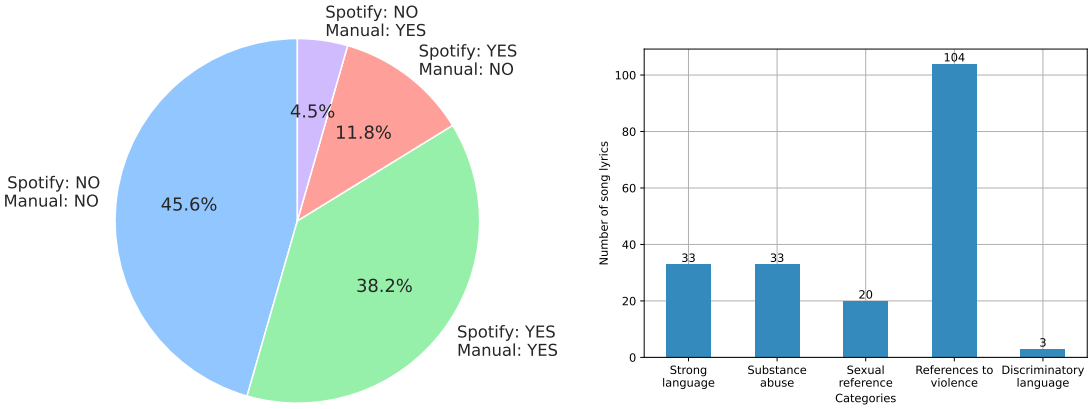

Figure 2. Partitioning of the song lyrics of the dataset according to the explicit (YES)/non-explicit (NO) information available on Spotify and our manual annotation (**left**); distribution of the reasons for the explicitness of their lyrics for the dataset songs that were "missing" the explicit tag on Spotify (**right**).

For 83.8% of the song lyrics of the dataset, the Spotify tags and our manual annotation of explicitness coincided: for 45.6% of the song lyrics, they both agreed that they do not contain explicit content, while for other 38.2% of cases, they both agreed that they contain content inappropriate for or harmful to children. For the remaining 16.2% of the song lyrics, we observed different information. First, 11.8% of the lyrics in the dataset (i.e., 471 lyrics) were marked explicit in Spotify, while they were not according to our manual annotation. The reasons for this disagreement are many:

- The song does not contain explicit content itself, but is contained in an album (or collection) that is marked as a whole as explicit;
- The song does not contain explicit content itself, but its cover art (or of the album or collection in which it is contained) contains explicit content;
- Different opinions (possibly due to different cultures, sensibilities, etc.) between the right holder providing the explicitness information to Spotify and our annotators.

While the latter is somehow inevitable when manual judgement is involved, the former two denote situations that go beyond the goal of this work (predicting explicitness based on the song *lyrics* alone). Yet, they point out an important aspect to consider when relying on data provided by online platforms for training and evaluating automatic systems for detecting explicit content in the words of the songs: the explicitness tag may be due to information that does not necessarily regard the lyrics of the song to which the tag is attached.

Figure 2 (left) also shows that, for 4.5% of the lyrics in the dataset (i.e., 178 lyrics), they were marked non-explicit in Spotify, while they contained content inappropriate for or harmful to children according to our annotators. Again, this may be due to different opinions between the right holder and our annotators, but it also points out the possible incompleteness of the information provided by online platforms with regard to the ex-

plicitness of the content they offer. Looking more in detail at this group of song lyrics, Figure 2 (right) shows the reasons for explicitness according to our manually annotated dataset: the most-frequent one was "Reference to violence", followed by "Strong language" and "Substance abuse", and then the others. Interestingly, the bar plot in Figure 2 (right), computed on the songs on Spotify that we argued were missing the explicit tag, is somehow different than the distribution computed on all the explicit lyrics (Figure 1 (left)): while this may deserve further investigation, which goes beyond the scope of this paper, it may also be a hint that, while some categories (e.g., "Strong language", "Sexual reference", "Discriminatory language") are better covered by copyright holders because they are probably easier to identify, others (e.g., "Reference to violence" and "Substance abuse") are less immediate and may be more affected by cultural, societal, or even personal biases.

5. Automatic Detection of Fine-Grained Explicit Lyrics

The main goal of the research behind this work was to understand the feasibility of developing a system for predicting the reasons for song explicitness based on the lyrics of the song. To address this goal, we developed a couple of multi-label classifiers based on classical and state-of-the-art technologies for text classification: (i) a one-dimensional convolutional neural network (1D CNN) exploiting FASTTEXT word embeddings (ii) and a sigmoid activated classifier built on top of DISTILBERT, a transformer-based language model.

5.1. Text Classifiers Compared

5.1.1. $\text{1D CNN}_{\text{FT}}$

A commonly used deep learning architecture for image, text, and audio classification is the Convolutional Neural Network (CNN) [14], a type of deep learning model that uses convolutions to identify and understand patterns in input data, capturing spatial and temporal dependencies. It is a model particularly suited for tasks involving bi-dimensional data, such as image classification and video analysis. A variation of the base model, called a 1D CNN, has been found to be effective for text classification tasks, such as sentiment analysis and question classification. A 1D CNN is composed of multiple layers of three types: *convolutional layers*, possibly arranged in a hierarchical structure and followed by a Rectified Linear Unit (ReLU) transformation, that are used to learn features from the input data thanks to a filter shifted over all of them; *pooling layers*, which are used to reduce the number of parameters in the input by applying an aggregation function thanks to a filter shifted across the entire input, so as to reduce complexity, improve efficiency, and prevent overfitting; and *fully connected layers*, which flatten the multiple feature vectors from the previous layers into a single vector for predicting the output labels, thus performing classification based on the features extracted by the previous layers and their different filters, via a softmax activation function. A 1D CNN is typically used with word vectors generated by unsupervised neural language models. Following the successful application in previous works for explicitness classification of song lyrics (e.g., [5,6]), we exploited FASTTEXT word embeddings [15] to feed the raw text to the network. We refer the reader interested in a more detailed description of CNNs for text classification to the comprehensive review by Minaee et al. ([16], Section 2.2), while the actual technical details of the $\text{1D CNN}_{\text{FT}}$ model developed for the considered task is detailed in the source code publicly released as part of the evaluation material.

5.1.2. DISTILBERT_{C}

Transformers are neural models that use an encoded–decoder architecture and an attention mechanism to understand relationships between input and output sequences. They have become the reference architecture for Natural Language Processing (NLP) and have enabled the development of large-scale, pre-trained deep language representations, known as Transformer-based Language Models (TLMs). These models can be fine-tuned for specific NLP tasks, such as text classification in our work, by adjusting the last output

layer of the transformer network. More specifically, when applying pre-trained language representations to downstream tasks, minimal task-specific parameters are introduced, and these parameters are trained on the specific task under consideration. The fine-tuning step basically involves updating the parameters of the pre-trained model to minimize a task-specific loss function. To perform multi-label text classification with TLMs, a sigmoid-activated classification layer is typically added on top of the pre-trained model. One of the first and most-popular TLMs is BERT [10], a deep bidirectional representation learned from unlabelled text collected from BooksCorpus and English Wikipedia, which achieved state-of-the-art performance on several NLP tasks when released. DISTILBERT [12] is a variant of BERT that uses knowledge distillation during pre-training to reduce the model size by 40% and inference time by 60% while still maintaining comparable performance to the full version of BERT. We refer the reader interested in a more detailed description of how TLMs can be used for text classification to the comprehensive review by Minaee et al. [16, Section 2.10], while the actual technical details of the DISTILBERT$_C$ model developed for the considered task is detailed in the source code publicly released as part of the evaluation material.

While it is beyond the scope of this work to identify the absolute best-performing system for the task, we decided to assess at least the performance of these two systems, following the findings regarding the automatic detection of explicit content in song lyrics reported in previous works (e.g., [6]):

- 1D CNN$_{\text{FT}}$ proved very effective and efficient for explicit lyrics' classification, outperforming other simpler approaches (e.g., logistic regression) and attaining a score on par with classifiers developed on top of TLMs while requiring much less computational power (training and testing can be performed on standard CPUs) than the latter (training and testing require substantial computation power as offered by GPUs);
- Among classifiers based on TLMs, DISTILBERT$_C$ achieved comparable scores to larger models (e.g., BERT) for explicit lyrics' classification, while being smaller and computationally less demanding, thus presenting itself as a good candidate for practical usages.

Both systems are developed in Python, leveraging existing relevant libraries (keras [17], FASTTEXT [15], hugginface transformers [18]).

5.2. Research Question and Evaluation Protocol

In this work, we addressed the following research question:

RQ Is it feasible to effectively determine the reasons for the explicitness of song lyrics via automatic text classification techniques?

To measure the effectiveness of the two considered systems, we leveraged the contributed dataset by comparing the "gold" annotation provided by the human annotators with the "predicted" labels suggested by the system. As the dataset was used also to train the considered systems, we adopted the well-known 10-fold cross-validation protocol, which is useful for obtaining an estimate of how well a system will perform on unseen data: the dataset is divided into 10 equal parts (folds), and each system is trained on 9 of the folds and tested on the remaining one. This process is repeated 10 times, with a different fold being used as the test set in each iteration. The results from each iteration are then averaged to provide an overall performance measure for the model.

At each iteration, we computed the typical classification performance scores (precision, recall, and F_1) on each of the categories considered (i.e., the five reasons for explicitness, in addition to the overall explicit/non-explicit tagging), basically treating each of them as a separate binary classification task: the song lyrics is (1) vs. is not (0) annotated with that category (i.e., explicit, strong language, substance abuse, sexual reference, reference to violence, discriminatory language). Specifically, for each classification class (0/1) of each category, we totalled: *True Positives* (TP^i), i.e., lyrics correctly predicted as in the class; *False Positives* (FP^i), i.e., lyrics predicted as in the class, but that belong to the other one; and *False*

Negatives (FN^i), i.e., lyrics that belong to the class, but are predicted as in the other one. We then computed:

- Precision $P^i = \frac{TP^i}{TP^i+FP^i}$: this measures how precise the method is on the class, independent of its coverage;
- Recall $R^i = \frac{TP^i}{TP^i+FN^i}$: this measures how extensively the class is covered by the method;
- $F_1^i = \frac{2 \cdot P^i \cdot R^i}{P^i+R^i}$: this combines the previous two in a single representative measure.

To combine the performance of the two classes, we averaged the corresponding metrics for the two classes, obtaining the so-called *macro-averaged metrics*: P, R, and F_1. The motivation for computing the scores on each class (0/1) of each category (e.g., "Strong language") and then averaging them per category is related to the unbalanced nature of the annotated data, especially for the reason of explicitness categories: e.g., there are many more song lyrics that do not contain "Discriminatory language" (3,853) than the ones that have it (147). Frequently, in classification tasks, scores are reported just for a single class, labelled as "positive" (e.g., the class corresponding to lyrics having "Discriminatory language"). However, in highly unbalanced scenarios, like the one considered, swapping the classes (i.e., labelling one or the other as positive) could yield a completely different classification performance. Therefore, we preferred to compute relevant scores on both classes independently and then derive aggregated scores by averaging them, a well-established practice in the literature (cf., [19]). Moreover, among possible strategies for averaging—macro, micro, and weighted—we opted for macro-averaged metrics as they are preferred in unbalanced settings over others as they better capture the performance of *both* classes. The Python scikit-learn library was used for computing all the metrics (https://scikit-learn.org/stable/modules/model_evaluation.html, accessed on 2 February 2023).

Following the findings of previous works (e.g., [6]), we configured the systems accordingly. For 1D CNN_{FT}, we set the following hyperparameters: *num_word* = 40,000, *max_sequence_length* = 750, *epoch* = 5, and *batch_size* = 128. For DistilBERT_C, a sigmoid-activated sequence classification layer was added on top of the pre-trained (uncased) version of DistilBERT and fine-tuned for the task, with a learning rate of 2×10^{-5} and with batches of 16 sequences, each having a maximum length of 512.

The experiments were conducted on the Google Colaboratory cloud platform. To run the 1D CNN_{FT}, a standard runtime using CPU-based processing was enough (2 Intel Xeon CPU @ 2.30 GHz were actually used), while for DistilBERT_C, a GPU-powered runtime was needed (an NVIDIA Tesla T4 GPU was actually used). All required Python modules for running the code for each classifier are detailed in the code itself, made available in the released evaluation material.

5.3. Evaluation Results and Discussion

Table 2 reports the performance of the two considered systems.

Both systems were very good at predicting the general explicitness of song lyrics scoring an F_1 of 0.802 (1D CNN_{FT}) and 0.862 (DistilBERT_C). The results, especially for DistilBERT_C, are in line with the ones reported in previous work ([6]), while slightly lower for 1D CNN_{FT}. We recall that the evaluation datasets were substantially different in the two works: in [6], (i) a larger dataset was used (800K vs. 4K lyrics), and (ii) the annotations were directly obtained from online platforms (cf., the discussion on the completeness and accuracy of the annotations in online platforms in Section 4).

Delving into the results related to the novel aspects investigated in this work, we observed dissimilarity between the systems in predicting the various correct reasons for explicitness. The scores of the two systems were very high and quite close at predicting song lyrics containing "Strong language", reaching an F_1 of 0.909 (1D CNN_{FT}) and 0.929 (DistilBERT_C). Instead, performances on the other categories were substantially different, with DistilBERT_C performing considerably better than the 1D CNN_{FT}: F_1 scores for DistilBERT_C ranged from a minimum 0.750 ("Substance abuse") to a max-

imum 0.886 ("Discriminative language"), while the ones for the 1D CNN_{FT} went from 0.483 ("Substance abuse") to 0.599 ("Discriminative language"). The actual F_1 difference between the two systems went from 0.210 ("Sexual reference") to 0.287 ("Discriminatory language"). Overall, DistilBERT_C outperformed the 1D CNN_{FT}, scoring $F_1 \geq 0.750$ for any category.

Table 2. Precision (P), recall (R), and F_1 scores of two multi-label text classification systems (1D CNN_{FT}, DistilBERT_C) for predicting the explicitness of song lyrics and the reasons for the explicitness. The highest scores are highlighted in bold.

Category	System	P	R	F_1
Explicit	1D CNN_{FT}	0.809	0.798	0.802
	DistilBERT_C	**0.861**	**0.862**	**0.862**
Strong language	1D CNN_{FT}	0.898	0.921	0.909
	DistilBERT_C	**0.906**	**0.941**	**0.922**
Substance abuse	1D CNN_{FT}	0.467	0.500	0.483
	DistilBERT_C	**0.745**	**0.755**	**0.750**
Sexual reference	1D CNN_{FT}	0.636	0.566	0.582
	DistilBERT_C	**0.820**	**0.771**	**0.792**
Reference to violence	1D CNN_{FT}	0.559	0.517	0.501
	DistilBERT_C	**0.765**	**0.788**	**0.775**
Discriminatory language	1D CNN_{FT}	0.684	0.571	0.599
	DistilBERT_C	**0.938**	**0.845**	**0.886**

Focusing on the results of DistilBERT_C, we observed that the two categories that were easier to predict were "Strong language" (0.922) and "Discriminative language" (0.886), while the toughest one was "Substance abuse" (0.750), followed closely by "Reference to violence" (0.775) and "Sexual reference" (0.792). A possible reason for this is that, for the former two, often explicitness was due to the usage of some well-recognized offensive words (i.e., easier to learn during training), while for the latter three, the explicitness may be more subtle and more related to the way the words are used, rather than the words per se.

Summing up, we believe the assessment, and in particular the performance scores of DistilBERT_C for each single category $F_1 \geq 0.750$, showed the feasibility of effectively tackling the problem of determining the reason(s) for the explicitness of song lyrics with text classification techniques, and thus, we can positively answer the investigated research question.

6. Limitations

In this section, we briefly summarize some possible limitations of the work. A first limitation concerns the manual annotation performed, which, as with any human activities based on subjective judgement, may be affected by possible personal opinions and cultural and demographic backgrounds. In our approach, we relied on two annotators in the process of the construction of the datasets, and although the measured agreement between the two was quite high (cf., the high score for the IAA reported in Section 4), involving other annotators in the construction or extension of the dataset may lead to minor differences in the resulting annotations. Moreover, we recall that a recent stream of works on data perspectivism (e.g., [20]) emphasizes the value of the disagreement between annotators, suggesting that all the annotations produced by the annotators, even the ones showing disagreement between them, should be included in the gold standard to avoid destroying any personal opinion, cultural influence, or rich linguistic knowledge as a result of the agreement and harmonization processes.

A second limitation concerns the language of the song lyrics considered in the dataset. All the work was conducted on English song lyrics; therefore, our findings can be considered representative at most for English-based songs. Enriching the dataset with song lyrics

in languages other than English may be useful to investigate possible differences across languages and cultural contexts.

A third limitation of the approach concerns the prediction performance reported in the evaluation. We recall that our goal was to investigate the research question, that is showing the feasibility of effectively tackling the problem of determining the reason(s) for the explicitness of song lyrics with text classification techniques. The performance of DISTILBERT$_C$ ($F_1 \geq 0.750$ for each single category considered) confirmed this. Although also other studies confirmed DISTILBERT$_C$ as one of the best-performing classifiers for assessing the explicitness/non-explicitness of song lyrics (cf., [6]), we are not claiming this is the absolute best scoring system for tackling the problem. That is, the evaluation results provide evidence of the performance a text classifier can achieve on the task, but the reported scores should not be considered as the best performance ML approaches can achieve for the investigated problem. Further experiments would be necessary to assess the best-scoring classifier for the task, involving, for instance, some of the 150 text classification methods discussed in [16].

7. Conclusions

In this paper, we investigated the yet-unexplored problem of predicting the type(s) of explicit content in song lyrics, that is assessing whether the song lyrics contain explicit content and, if so, determining if it is due to strong language, sexual references, and so on.

To tackle the problem, we first identified five main reasons for considering song lyrics as explicit and contributed a novel, publicly released dataset of 4000 song lyrics, manually annotated according to these five reasons. We provided a detailed analysis of the dataset, comparing also the manual annotations with information on explicit content made available by online streaming platforms (namely, Spotify), showing that differences do exist.

Leveraging the contributed dataset, we then trained and evaluated a couple of text classification systems to assess the feasibility of tacking the problem with machine-learning-based techniques: one based on a one-dimensional convolutional neural network and FASTTEXT word embeddings (1D CNN$_{\text{FT}}$) and one obtained by adding a sigmoid-activated sequence classification layer on top of a pre-trained language model (DISTILBERT$_C$). Both systems performed extremely well at predicting (just) the explicitness of the song lyrics' content, but DISTILBERT$_C$ substantially outperformed the 1D CNN$_{\text{FT}}$ in determining the reasons for the explicitness. The performance scores varied on the different reasons of explicitness, but DISTILBERT$_C$ scored $F_1 \geq 0.750$ for each of them, thus confirming the feasibility of predicting the type(s) of the explicit content of song lyrics via text classification. As a further contribution of the work, a fine-tuned version of DISTILBERT$_C$ for predicting the type(s) of explicit content in song lyrics has been publicly released.

Future work will focus on three main directions: (i) enlarging the dataset to have more material to train and evaluate systems for tackling the task; (ii) comparing multiple state-of-the-art multi-label text classifiers, in order to identify the best-performing approach for the specific scenario of predicting the type(s) of explicit content in song lyrics; and (iii) investigating the applicability of recent prompt-based approaches (e.g., InstructGPT/ChatGPT [21]) for the task, studying, in particular, the more effective way to formulate the request.

Author Contributions: Conceptualization, M.R.; methodology, M.R.; software, M.R.; validation, M.R.; resources, M.R. and S.E.; data curation, M.R. and S.E.; writing—original draft preparation, M.R.; writing—review and editing, M.R. and S.E.; supervision, M.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Informed Consent Statement: Not applicable.

Data Availability Statement: The novel dataset developed in this work, a trained model for predicting the reasons for song lyrics explicitness, and the evaluation material are made available at https://github.com/rospocher/explicit-lyrics-detection (accessed on 2 February 2023).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Chin, H.; Kim, J.; Kim, Y.; Shin, J.; Yi, M.Y. Explicit Content Detection in Music Lyrics Using Machine Learning. In Proceedings of the 2018 IEEE International Conference on Big Data and Smart Computing, BigComp 2018, Shanghai, China, 15–17 January 2018. IEEE Computer Society: Washington, DC, USA, 2018; pp. 517–521. [CrossRef]
2. Kim, J.; Yi, M.Y. A Hybrid Modeling Approach for an Automated Lyrics-Rating System for Adolescents. In Proceedings of the Advances in Information Retrieval—41st European Conference on IR Research, ECIR 2019, Cologne, Germany, 14–18 April 2019; Proceedings, Part I; Azzopardi, L., Stein, B., Fuhr, N., Mayr, P., Hauff, C., Hiemstra, D., Eds.; Springer: Berlin/Heidelberg, Germany, 2019; Volume 11437, pp. 779–786. [CrossRef]
3. Bergelid, L. Classification of Explicit Music Content Using Lyrics and Music Metadata. Master's Thesis, KTH, School of Electrical Engineering and Computer Science (EECS), Stockholm, Sweden, 2018.
4. Fell, M.; Cabrio, E.; Korfed, E.; Buffa, M.; Gandon, F. 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, Marseille, France, 11–16 May 2020; European Language Resources Association: Marseille, France, 2020; pp. 2138–2147.
5. Rospocher, M. Explicit song lyrics detection with subword-enriched word embeddings. *Expert Syst. Appl.* **2021**, *163*, 113749. [CrossRef]
6. Rospocher, M. On exploiting transformers for detecting explicit song lyrics. *Entertain. Comput.* **2022**, *43*, 100508. [CrossRef]
7. Rospocher, M. Detecting explicit lyrics: A case study in Italian music. *Lang. Resour. Eval.* **2022**. [CrossRef]
8. Vaglio, A.; Hennequin, R.; Moussallam, M.; Richard, G.; d'Alché-Buc, F. Audio-Based Detection of Explicit Content in Music. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 526–530.
9. Fell, M.; Cabrio, E.; Corazza, M.; Gandon, F. Comparing Automated Methods to Detect Explicit Content in Song Lyrics. In Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019), Varna, Bulgaria, 2–4 September 2019; INCOMA Ltd.: Varna, Bulgaria, 2019; pp. 338–344. [CrossRef]
10. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. 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), Minneapolis, MN, USA, 2–7 June 2019; Association for Computational Linguistics: Minneapolis, MN, USA, 2019; pp. 4171–4186. [CrossRef]
11. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv* **2019**, arXiv:1907.11692.
12. Sanh, V.; Debut, L.; Chaumond, J.; Wolf, T. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *arXiv* **2019**, arXiv:1910.01108.
13. Passonneau, R. Measuring Agreement on Set-valued Items (MASI) for Semantic and Pragmatic Annotation. In Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06), Genoa, Italy, 22–28 May 2006; European Language Resources Association (ELRA): Genoa, Italy, 2006.
14. LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [CrossRef]
15. Bojanowski, P.; Grave, E.; Joulin, A.; Mikolov, T. Enriching Word Vectors with Subword Information. *Trans. Assoc. Comput. Linguist.* **2017**, *5*, 135–146. [CrossRef]
16. Minaee, S.; Kalchbrenner, N.; Cambria, E.; Nikzad, N.; Chenaghlu, M.; Gao, J. Deep Learning–Based Text Classification: A Comprehensive Review. *ACM Comput. Surv.* **2021**, *54*, 1–40. [CrossRef]
17. Chollet, F. Keras. 2015. Available online: https://keras.io (accessed on 2 February 2023).
18. Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; et al. Transformers: State-of-the-Art Natural Language Processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, online, 16–20 November 2020; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 38–45.
19. Chicco, D.; Jurman, G. The advantages of the Matthews correlation coefficient (MCC), over F1 score and accuracy in binary classification evaluation. *BMC Genom.* **2020**, *21*, 6. [CrossRef] [PubMed]
20. Basile, V.; Cabitza, F.; Campagner, A.; Fell, M. Toward a Perspectivist Turn in Ground Truthing for Predictive Computing. *arXiv* **2021**, arXiv:2109.04270.
21. Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.L.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. Training language models to follow instructions with human feedback. *arXiv* **2022**, arXiv:2203.02155.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.