POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

Hit Song Prediction system based on audio and lyrics embeddings

Laurea Magistrale in Music and Acoustic Engineering

Author: Elisa Castelli

Advisor: Prof. Massimiliano Zanoni

Academic year: 2022-2023

1. Introduction

Music industry market is drastically changed with the diffusion of digital audio formats and the growth of streaming platforms. Technologies used in these last 15 years, to manage and analyze songs data have evolved. In particular, the development of machine learning has given rise to Music Information Retrieval (MIR), a field that employs computational techniques based on machine learning and deep learning to retrieve musical information and analyze them. MIR finds many applications such as music classification problems, recommendation systems and many others. Among all these applications, Hit Song Science (HSS) aims to investigate whether a song has the potential to become popular or not. Producing an artist or deciding to invest in marketing campaign for a song requires to consider several factors to understand the risks and the possible results in terms of popularity and profits. In this scenario technology innovation might play an important role. Nowadays, HSS is an hot and active research topic in MIR. It's important to precise that HSS's main goal is not to substitute talent scouts but, given the great amount of new songs released every day, it can be a useful tool to make a preliminary automatic selection of songs that can be appealing from an artistic or a market perspective.

Researches conducted in these last years, leading to the state-of-the-art in the field, propose Hit Song Prediction systems that employ mainly machine learning methods to address this challenge, as in [5], while only in few cases deep learning. Even exploiting deep learning approaches, systems proposed are based on Multi-Layer Perceptron (MLP) structures applied on songs metadata and sets of manually hand-crafted features, computed from audio and lyrics.

Starting from these observations, encouraged by the advancements in deep learning technologies able to automatically extract features from audio and texts, this thesis aims to address the Music Popularity Prediction problem, specifically predicting whether a song can become a hit or not based on its objective characteristics. This involves assigning popularity classes and predicting popularity scores, essentially turning the problem into either a multi-class or regression task. The novelty of this thesis lies in the exploration of embeddings power in this domain, seeking to leverage them as input features for the prediction model. In fact, the input feature vector used to compute the prediction result combines three components: audio embeddings computed from song mel-spectrogram using a ResNet-50 [1] model, text embeddings extracted with Sentence-BERT [4] from song lyrics and the

song release year.

2. Problem Formulation

Hit Song Prediction aims at computing the popularity of a song taking as input some features and metadata relative to the track examined. The prediction problem can be faced as a regression or a classification task.
In this thesis, the HSP problem is tackle with a novel approach based on deep learning methods, in particular exploiting embeddings extraction from audio and text, to build a feature vector for each song, obtained concatenating this two contributions and the relative release year. These design choices derive from some observations on the state of the art:

- A deep learning approach is not yet applied at its full potential in this field.
- Due to a lack of datasets of considerable size for HSP task and moved by the positive results obtained in researches, we want to employ audio and text embeddings taking advantage of transfer learning.
- The majority of the papers, for example [2], uses as features some biasing information, such as the artists popularity. This choice means that songs of already famous artists are facilitated to obtain high popularity score. Instead, we want to conduct our study starting from raw data such as the audio file content and the lyrics transcription of each song.
- Many methods proposed in HSP do not consider the release year of songs, even if researches demonstrate how the presence of this feature can have benefits on the final results adding a temporal context to songs.

3. Proposed Method

The model we propose for approaching the HSP problem is shown in a schematic representation in Figure 1. The system takes in input the .mp3 audio file, the song's lyrics and the song's release year. Then the workflow is divided in two autonomous branches: an audio and a lyrics processing chains.
These two branches process in an independent way the audio mel-spectrogram and the lyrics to obtain the correspondent embeddings. After the application of these pipelines, the two contributions are concatenated by adding the release year. The feature array obtained is given in input to a final multi-layer perceptron that returns an output of two types: the expected popularity score, between 0 and 100, or a popularity class, computed as follows:

- Low Popularity (class 0): popularity value between 0 and 24
- Mid-Low Popularity (class 1): popularity value between 25 and 49
- Mid-High Popularity (class 2): popularity value between 50 and 74
- High Popularity (class 3): popularity value between 75 and 100

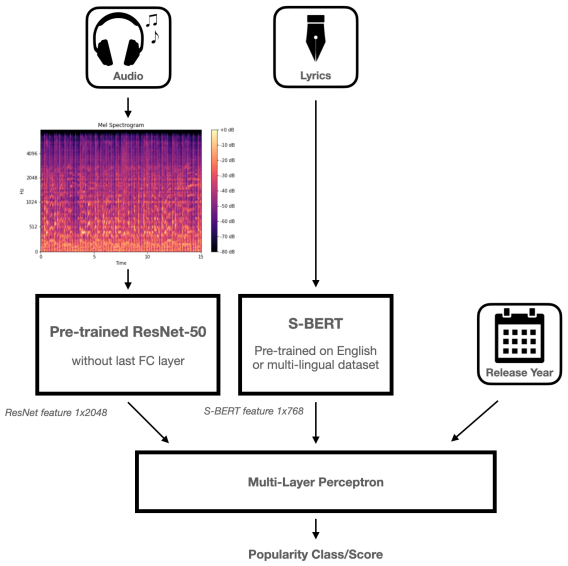

Figure 1: Overall system

3.1. Audio Processing Chain

Audio pipeline has the goal of producing audio embeddings starting from an audio .mp3 file that contains 30 seconds of the song. In order to achieve this task, first of all the mel-spectrogram of the audio is computed. Then the obtained 2D representation of the track, after a normalization process, is passed to a CNN-based neural network to extract the audio embeddings. In particular, we use a ResNet-50 [1] model pretrained on GTZAN Genre Dataset, due to its ability of automatically filtering and capturing relevant features of images.
After pre-training the model, it is used to perform transfer learning and to return embeddings from the music tracks of which we want to predict the popularity. In order to extract the embeddings, the model is used with all its layers except the last fully connected one. In this way,

the result obtained given a mel-spectrogram is a feature vector of dimension 1x2048.

3.2. Lyrics Processing Chain

Regarding the text embedding extraction, after retrieving each song lyrics, it is necessary to perform text vectorization, to achieve a compact numerical representation that preserve the meaning of the text and that can be easily processed by the system. Our system is working with lyrics, hence the focus is considering the meaning of texts. For this reason we develop a text processing chain that takes advantage of a Sentence-BERT [4] transformer that is also able to produce an homogeneous representation in size, independently from the length of the text. The input lyrics are pre-processed, with tokenization, lowercasing and any other necessary text cleaning steps. Successively, the sentences pass through a pre-trained sentence embedding model, which is capable of producing contextual embeddings for each token in the sentences. Finally, an average pooling strategy is applied to obtain a fixed-size sentence embedding. S-BERT model is developed for various languages and correspondent different pre-trained version of it are available. We evaluate our model on two different datasets: one that contains solely English lyrics and a more extended version of it that also contains multi-lingual songs. According to this choice, we use two version of S-BERT: *all-mpnet-base-v2* for the English dataset and *multi-qa-mpnet-base-dot-v1* for the multi-lingual one. Both the models produce an output lyrics embedding of size 1x768.

3.3. Combination of the two systems

These two kind of embeddings are then concatenated with the release year x_{year} to create the final feature vector that represent each song x. The aim of this final part of the system is to take all these contributions and compute the popularity prediction as:

$$P_x = f(e_a, e_t, x_{year}) \ , \tag{1}$$

having as f the neural network function that computes the popularity, e_a and e_t respectively the audio and the text embeddings.
The Multi-Layer Perceptron in charge to process these information has different technical settings based on the problem addressed. In fact, different version of MLP are implemented based on the dataset used and the experiment conducted, as reported in 4.2. In general, every model used during the experiments have a structure that can be resumed as: input layer, some hidden layers and output layer. Each layer uses Batch Normalization, linear layer with ReLU activation function and Dropout. As result of the output layer, based on the task the model has to deal with, the MLP returns the popularity class assigned to the song given as input or the popularity score.

4. Experimental Setup and Evaluation Methods

In this section, we will explain the system setup used to conduct the experiments and investigate the effective ability of our system in predicting song popularity. Before that, the dataset creation process is depicted. At the end results obtained during the tests are reported.

4.1. Dataset

Starting from the SPD dataset proposed by [2] in 2019, some evidences emerged proving the presence of corrupted data. For this reason a cleaning process is conducted, following the steps illustrated in Figure 2.

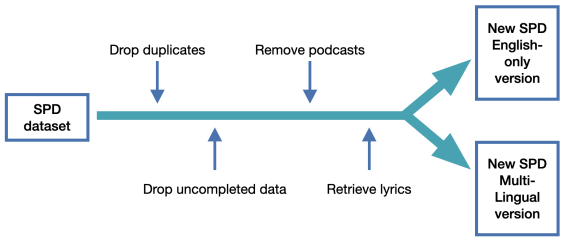

Figure 2: Dataset creation process

After having drop duplicates and uncompleted annotated data, we have to discriminate podcasts and songs to only maintain the latter ones. Finally, a mismatch between titles and lyrics of songs has been noticed, hence updated lyrics are retrieved using MusixMatch API. Initially, we collect only the English songs then, to increase the size of the data, we create also a multi-lingual version of the dataset considering also Italian, German, French, Spanish and Portuguese songs. The final English dataset, called *SPD-English*, has 26711 songs while the multi-

lingual one, called *SPD-Multilingual*, has 41333 songs. Both of them have a Gaussian-shaped popularity distribution.

4.2. Training Setup

The training setup involves two main sub-tasks: the pre-training of the ResNet-50 model in charge to extract audio embeddings and the training process of the final Multi-Layer Perceptron that computes the popularity prediction.
Regarding the **ResNet-50 pre-training** procedure, we fine tune the original ResNet-50 pre-trained on ImageNet on another dataset related to music, with the aim of adapting the model to the specific characteristics and patterns present in audio data. In particular, the task on which we pre-train the model is genre classification, using the dataset GTZAN Genre Dataset. In order to increase the size of the dataset and make the model more robust and less sensitive to overfit we use some data augmentation techniques: time stretching, pitch shifting and SpecAugment. Each one of this methods contributes to change some characteristic of the audio, maintaining the others intact, creating at each training iteration two variations of the same signal.
In order for the ResNet-50 to be trained on the GTZAN genre classification task, we change the number of output neurons of model's last layer to 10, that is the number of classes represented in the dataset. Before this last layer we add a Dropout layer with probability p = 0.5 and a Batch Normalization layer. Other hyper parameters used in the pre-training process are: the learning rate to 1e-5, batch size to 16, optimizer of type Adam with weight decay 1e-1.
For what concern the **Multi-Layer Perceptron training** phase, we can distinguish some shared parameters between all the model configuration used during the tests and other characteristics that are specific for each model.
First of all, based on the dataset used, the problem addressed and the test in which they are employed we can distinguish four model setups:

- **Audio-only setup**, with SPD-English dataset, neglecting lyrics to perform classification
- **Audio and lyrics basic setup**, with SPD-English dataset considering one single lyrics embedding to perform classification
- **Audio and lyrics English weighted setup**, with SPD-English dataset and double text embeddings to perform classification and regression
- **Audio and lyrics multi-lingual weighted setup**, with SPD-Multilingual dataset and double text embeddings to perform classification and regression

These four configurations can be summarized in two principal setups, based on the fact that they use the English or the multi-lingual dataset. For every training procedure, data augmentation is applied, following the same approach already explained for the ResNet-50 pre-training process. In every experimental setup, performing classification task over the four classes of popularity, the loss function used is a *Cross Entropy Loss*, with mean reduction. In regression case instead the loss function used is the *Mean Absolute Error*. As optimizer we have chosen Adam with *learning rate* 1e-5 and *weight decay* 1e-2.
Regarding all the **English models** that use that SPD-English dataset, too complex architectures lead the model to overfit. For this reason, the MLP structure chosen is composed by the input layer, one hidden layer and the output layer. The input layer dimension changes according to the particular configuration used:

- Audio and lyrics English weighted setup needs $n = 3585$ input features. Since it uses a weighted version of the embedding, the input features dimension is obtained by summing the audio embedding of size 1x2048, two times the lyrics embedding of size 1x768 and the release year.
- Audio and lyrics basic setup needs $n = 2817$ input features because it employs the audio embedding, one single text embedding and the release year.
- Audio-only setup needs $n = 2049$ input features becasue it takes only the audio embedding and the release year as information to process.

In each of these setups, the input layer maps the features into 512 neurons. The hidden layer receives the data and maps them into other 128 neurons. The final layer takes the 128 features and maps them to four neurons in case of classification, otherwise in a single neuron if performing regression. At each step, linear layer with ReLU activation function is used while regularization

and dropout are applied to reduce overfitting.
Considering the **Multi-lingual models**, that use SPD-Multilingual dataset, the MLP consists in one input layer that takes $n = 3585$ input features and maps them into 1024 neurons. Then three hidden layers uses 1024, 512 and 128 neurons to process features and a final layer performs classification or regression, using four or one neuron accordingly to the task addressed.
Also in this setup, the layers are implemented using Linear layer with ReLU activation function. In case of classification, each one of the four neurons represents a popularity class in which the MLP can map a song. On the other hand, from a regression perspective, the final neuron uses a Sigmoid activation function.

5. Results

Different tests are conducted to evaluate the performance of the system and investigate the effectively applicability of our solution in Hit Song Prediction problem. First of all, to investigate the influence of lyrics embedding in Hit Song Prediction, we conduct a comparison between the performances of an audio-only based model with respect to the performances of the same model that also takes in input the lyrics.

Model	Dataset	Training Accuracy	Validation Accuracy
Audio-Only	SPD-English	0.532	0.565
Audio and Lyrics	**SPD-English**	**0.577**	**0.595**

Table 1: Comparison among the performance obtained in popularity class prediction considering only audio embeddings and using both audio and text embeddings

As can be seen in Table 1, results align with the literature, because they confirm that a multimodal approach, that captures both the musical and lyrical aspects of a song, leads to better performance in music related tasks, in particular in music popularity prediction. Furthermore, we demonstrate the effectiveness of Sentence-BERT in capturing semantic characteristic of texts useful for HSP that is a field in which it has never been used. The demonstrated utility of lyrics in HSP reveals that the subject of lyrics can influence a song popularity, reflecting cultural trends or the spirit of a particular time.
Having proved that text embeddings, computed from lyrics, bring to an improvement in the system performance, we investigate the possibility of weighting embeddings.
In fact, after performing different attempts, changing the MLP input vector constitution, we decide to use in the next tests the same text embedding twice, with the aim of propagate the textual information through multiple layers, providing more opportunities for the model to learn complex patterns and relationships within the text data.
After this first step, we move our attention to the system overall results, in two main case scenarios: popularity score and popularity class prediction. Doing this evaluation of classification and regression capabilities of our system, we use as reference system the architecture proposed by Pham [3] and the one implemented by Gutierrez [2] that is, in our knowledge, the state-of-the-art in HSP, using deep learning methods.

Model	Study	Dataset	N° of classes	Validation Accuracy (%)
MLP	[3]	MSD	2	79.3
HitMusicNet	**[2]**	**SPD**	**3**	**83.03**
Our model	Proposed	SPD-English	4	67.95
Our model	Proposed	SPD-Multilingual	4	70.14

Table 2: Comparison among the performance in popularity classification considering existing models and our proposed approaches

Analyzing the results reported in Table 2, we can see that the multi-lingual method proposed obtains better results in terms of accuracy with respect to the English-only version. This demonstrates that greater amount of songs influences the results, bridging the performance gap of multi-lingual Sentence-BERT with respect to the one trained only in English. Moreover, we can evaluate our results comparable with the ones obatined by [2, 3] if we consider the situation in which we set ourselves, with the constraints of:

- Augmenting the number of output classes, with respect to the other models, that increases the complexity of the problem.
- Discarding any metadata, such as the artist's popularity that [3] identify as one of the most influential features in their dataset. Taking in input only audio preview file, lyrics and release year of each song, we keep our solution as unconditioned as possible from data related to popularity that can affect the prediction.

This result leads us to claim that effectively the

combination of audio embeddings, text embeddings and release year in music popularity prediction is a valid approach even if it has different grades of improvement and the resulting accuracy can surely enhance with more data at disposal.

Model	Study	Dataset	MAE	MSE
MLR	[3]	MSD	0.1357	0.0184
MLR + Lasso	[3]	MSD	0.1342	0.0180
HitMusicNet	**[2]**	**SPD**	**0.0855**	**0.0118**
Our model	Proposed	SPD-English	0.1026	0.0167
Our model	Proposed	SPD-Multilingual	0.0937	0.0145

Table 3: Comparison among the performance in popularity regression considering existing models and our proposed approaches

Also performing regression, we compare the results obtained by our model both using English and multi-lingual datasets. The results of the experimental models in terms of Mean Squared Error and Mean Absolute Error, presented in Table 3, show that the model proposed trained using the multi-lingual dataset outperforms the one trained using only English songs. Even in this case, we can see how the number of songs affects the overall performance. Moreover, the results obtained by both our proposed methods outperforms the MAE and MSE value achieved by [3] and are close to the ones obtained by [2]. In this case, there is no mismatch in the number of output prediction classes and we are always neglecting artist popularity information. Looking at the accuracy values reached it is more clear that our results are comparable with the state-of-the-art. For this reason we claim the potential of our system, that is based mainly on audio and text embeddings, to tackle HSP, even if there's surely room for improvement.

6. Conclusions

In this work we have proposed an Hit Song Prediction system, based on audio and lyrics embeddings. As our knowledge this is the first work that aims to address this problem using multimodal automatically extracted embeddings, instead of employ hand-crafted features. Results evidence positively how this method can be considered effective for Hit Song Prediction, although many improvements and further develops can be introduced. First of all, a different ResNet-50 pre-training procedure could be followed, using a dataset that contains more data with more target classes to extract meaningful embeddings from the audio mel-spectrograms and enhance the ability of the system of capturing relevant aspect of audio. Moreover, different kind of CNN-based methods can be evaluated, for example the Audio Spectrogram Transformer. Another improvement could be using a greater dataset for evaluating songs popularity. Starting from Spotify's and MusixMatch's API a dataset specifically designed for this task could be created collecting a more relevant amount of songs with updated popularity scores.
Thinking about the applications of our model, could be interesting in future to use it in a reverse way with the aim of answering the question "Which are the characteristics of a song that makes it popular?". In order to investigate which are the most important features of audio and lyrics that mostly influence and determine how much a song is appealing for audience.

References

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[2] D. Martín-Gutiérrez, G. Hernández Peñaloza, A. Belmonte-Hernández, and F. Álvarez García. A multimodal end-to-end deep learning architecture for music popularity prediction. *IEEE Access*, pages 39361–39374, 2020.

[3] James Q. Pham. Predicting song popularity. 2015.

[4] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019.

[5] Mengyisong Zhao, Morgan Harvey, David Cameron, Frank Hopfgartner, and Valerie J. Gillet. An analysis of classification approaches for hit song prediction using engineered metadata features with lyrics and audio features. *iConference*, 2023.