Unsupervised Heterogeneous Graph Neural Network for Hit Song Prediction through One Class Learning

Angelo Cesar Mendes da Silva and Marcos Paulo Silva Gôlo and Ricardo Marcondes Marcacini

Instituto de Ciências Matemáticas e de Computação - Universidade de São Paulo (USP)
{angelo.mendes, marcosgolo, ricardo.marcacini}@usp.br

Abstract. Although the concept of success is subjective, it can be related to the popularity and interest of users. Measuring the success of a song in advance allows for offering information of great interest to the music market. Hit song prediction is an existing task in Music Information Retrieval that explores approaches for measuring music success based on features. Musical data is intrinsically multimodal, where features from different sources have complementary semantic information. Therefore, structuring musical data and building a unique space that embeds multiple features is a challenge in musical data representation. Using heterogeneous graphs to structure multimodal data is a resource for exploring the intrinsic semantic relationship between features. In this sense, this work proposes to structure musical features over heterogeneous graphs and learn a new graph-based multimodal representation for songs using an unsupervised graph neural network to handle the hit song prediction task. We formulated the hit song prediction task as a one-class learning problem to mitigate the non-hit song gaps and highlight the hit song as the interest class. We measure the performance of representations based on lyrics and artist features and present promising results using our learned representations that outperform other strategies for representing musical data.

CCS Concepts: • **Computing methodologies** → **Artificial intelligence**; **Machine learning algorithms**; *Neural networks*.

Keywords: graph-based representation, heterogeneous graph, music representation, one class hit song prediction

1. INTRODUCTION

In recent years, the Music Information Retrieval (MIR) literature has highlighted the changes in consuming music with the popularization of streaming platforms. The impact of this popularization reflects how artists produce music [Hiller and Walter 2017]. Similar to social media content production strategies, it has been shared for artists to release multiple single songs rather than build albums [Krohn-Grimberghe 2021], creating new challenging scenarios for music information retrieval.

An example is the hit song prediction task, which aims to estimate the song's success. The success information is of great interest to the music market and is reported annually [IFPI 2022]. Generally, the success of a song is associated with rankings that can be defined by technical criteria or based on the song's popularity obtained through users' consumption [Song et al. 2012]. Therefore, a relevant hit song prediction challenge is defining which features must be present in the songs to become successful.

Music data is defined as multimodal data, whose structure is composed of features of different types, such as audio and text modalities. In MIR, the explored task is decisive in choosing the feature set used to represent the music [Kim et al. 2020; Simonetta et al. 2019]. The absence of music datasets with multimodal features is a problem for handling multimodal representations in MIR tasks [Karydis et al. 2018; Chen et al. 2019]. The absence of music datasets with multimodal features is a drawback

This work was partially funded by CAPES grant number [88887.671481/2022-00, 12049601/D], and the Center for Artificial Intelligence (C4AI-USP) and the support from the São Paulo Research Foundation (FAPESP) grant number [2019/07665-4].

for handling multimodal representations in MIR tasks, being necessary to build more robust datasets.

Representation learning is an approach to building feature vectors that can be formed by multimodal information without dependence on a target task, reducing the need for manual pre-processing and hand-crafted feature extraction of the original data [Bengio et al. 2013]. Moreover, graph-based musical representations have obtained state-of-the-art results in several MIR tasks by structuring data over graphs and building multimodal representations that explore the existing relations between features from different modalities [Xia et al. 2021; Shi 2022].

Generally, the hit song prediction problem is solved by binary learning, where the objective is to predict whether a song will be a hit or non-hit by learning from past hit and non-hit songs [Herremans et al. 2014; Pareek et al. 2022]. However, the scope of non-hit songs is wide since exists more non-hit songs than hits (imbalanced scenario), so it's hard to cover that scope and label non-hit songs. Furthermore, the interest is only in hits songs. Therefore, binary algorithms need non-hit songs having no interest in them. An alternative is One-Class Learning (OCL) which learns only with an interest class (e.g., hit song) and predicts if the example will be of the interest class or not [Tax 2001; Emmert-Streib and Dehmer 2022]. OCL can mitigate the limitations of binary learning since in the OCL: (i) we do not cover the wide non-hit songs scope; (ii) we focus only on hit songs (iii) we better support the imbalanced scenario since we use only one class; and (iv)we do not label non-hit songs.

This work uses a dataset with hit songs from Spotify between 2000 and 2019. We add the lyric and artist relations information to enrich the dataset. We propose and evaluate an approach to structure the music data through a heterogeneous graph and use an Unsupervised Graph Convolutional Network (GCN) to embed artist and music features in a unified space representing the music. To predict hit songs, we define a criterion to label hit songs similar to the Spotify criterion and use OCL to estimate the music's success through classification and regression. Our main contributions are three-fold:

(1) We model a heterogeneous graph with artist and song nodes to enrich the song representation for hit song prediction;
(2) We propose an unsupervised GCN to learn the song embeddings for hit song prediction;
(3) We formulate the hit song prediction problem as a one-class learning problem.

Experimental results indicate that our approach achieves competitive results, even compared to state-of-the-art models based on deep neural language models (BERT) to represent artists' lyrics and textual information. Moreover, our method based on one-class learning is more suitable for the hit song prediction task, being more natural to consider only historical hit songs for model training.

2. BACKGROUND AND RELATED WORK

Hit song prediction. Hit song prediction is currently a relevant task explored in MIR [Bertoni et al. 2021]. Generally, the approaches handle the hit song problem as a binary learning problem [Herremans et al. 2014; Singhi 2015]. This strategy needs annotated instances for hit and non-hit classes. The musical representation used to define a hit song has different features. We noted works such as [Pareek et al. 2022] that use previously processed high-level audio features, [Zangerle et al. 2019] that utilized low-level audio features as a base to learn a deep representation, or some approaches that use pre-processing resources as Principal Component Analysis [Ge et al. 2020] or autoencoders [Martín-Gutiérrez et al. 2020]. The lyrics concentrate on important musical information and are also explored in [Singhi and Brown 2015; Martín-Gutiérrez et al. 2020]. On the other hand, we learn a new representation for musical data that embed lyrics and artist features. Finally, we formulate the task as one class prediction problem.

Graph-based representation. Modeling data through graphs is a resource that allows embedding multiple modalities from unique data in a single structure and embedding information shared between

related graph objects [Yang et al. 2020; Wu et al. 2021]. The mapping of unstructured data through graphs to represent the data has been introduced in several applications and presented state-of-the-art results[Ali and Melton 2019; Sawant and Prabukumar 2020; Xia et al. 2021]. In MIR, the graph modeling for representation learning was successful in artist similarity [Korzeniowski et al. 2021] and emotion recognition [Silva et al. 2022], but there is a gap in many tasks, such as hit song prediction. We structured artist and music features through a heterogeneous graph and explored object relations in an unsupervised way to learn a new graph-based multimodal representation for music data. We do not use the hit and non-hit labels since this allows us to reproduce the process in other MIR tasks.

One class learning. OCL is applied when we are interested in one class, e.g., hit songs, without relying on data from other classes, e.g., non-hit songs. OCL is commonly exploited in binary problems, for instance, in fake news detection, relevant app review classification, and interest event detection. OCL is relevant for the hit song prediction task because, considering the real music market scenario, we have few hit songs in relation to non-hit. This problem has been highlighted in MIR [Ge et al. 2020; Martín-Gutiérrez et al. 2020; Bertoni et al. 2021; Pareek et al. 2022] for hit-song prediction. However, it is explored as a problem with a dependency of non-hit songs.

In summary, our work focuses on the gaps for graph-based multimodal learning to represent musical data with application in the hit song prediction task through one-class learning. We explore features of heterogeneous graphs to model a framework for music data using artist and music features and apply it as input for an unsupervised GCN to learn a multimodal representation for the data. Thus, we expect a representation with more semantic information and greater discriminative power for the one-class hit song prediction that does not need non-hits to classify a new hit song.

3. PROPOSED APPROACH

We aim to use content-based musical features to learn a data representation to classify the song as a hit song or non-hit song. We formulate the hit song prediction task as a one-class learning problem $Y = f(X)$, where Y represents the hit song label, X represents the song's features, and $f(.)$ denotes the function responsible for classifying the song as a hit from a multimodal musical representation as input. We formulate this task as one class problem because, in this scenario, we need only instances of interest class to train the predictive models. Therefore, our model trains only hit song instances to predict the success of new songs.

The proposed method for handling the hit song prediction is divided into three steps. Initially, we structure the musical data about a heterogeneous graph formed by two types of nodes with artist and song features and relationships between them. Then, in sequence, we introduce an unsupervised heterogeneous graph convolution network to embed the multimodal features of nodes into a single feature space that represents the songs. Finally, we annotate these song embeddings according to the popularity label, and we use a one-class classifier to predict the hit song instances.

3.1 Heterogeneous graph modeling for musical data

Musical perception is intrinsically multimodal and has complementary information [Knees and Schedl 2013]. Users who listen to music absorb multiple pieces of information as audio through the beat while they identify the artist and lyrics. We note different approaches to model music data through heterogeneous graphs related to the hit song prediction. [Melo et al. 2020] explores the genre classification task in which the heterogeneous graph has nodes related by a similarity metric between the music features. [da Silva et al. 2021; da Silva et al. 2022] explores the tasks of instance selection and music retrieval and adopts the same strategy based on clustering information to relate the different nodes. [Korzeniowski et al. 2021] aims to predict the similarity between artists and related nodes with previous annotations. In our heterogeneous graph modeling, we have artist and song types. The songs and artists have a direct relation, while pre-annotated data give the relations between artist and

artist. We detail the music data in Section 4.1. Figure 1 illustrates the heterogeneous graph built.

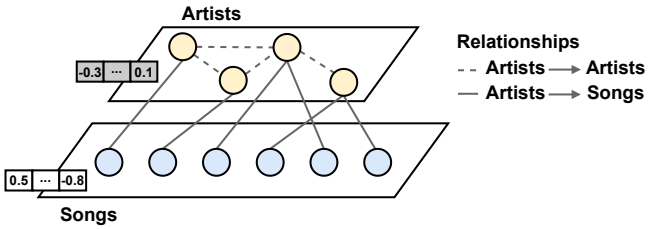

Fig. 1. Our heterogeneous graph proposed to structure musical data for the hit song prediction. We have artists and songs as nodes and the relationships between artists and artists; and artists and songs. All nodes have initial features.

Formally, our heterogeneous network is defined as $N = (O, R, W)$, where O represents the set of objects, R represents the relationships, and W represents the weights. Let $o_i \in O$ as the notation for an object, $r_{o_i,o_j} = (o_i, o_j) \in R$ indicates whether there is a connection between the objects o_i and o_j, where the weight of r_{o_i,o_j} is given by w_{o_i,o_j} with $w \in W$. Currently, our proposal uses binary relations, $w_{o_i,o_j} \in \{0, 1\}$, for instance, if a song node is associated with an artist node or an artist node is associated with another artist node. The benchmark dataset used has textual and acoustic features so that the set of objects $O = \{O_S \cup O_A\}$ organizes each feature modality in the heterogeneous network proposed for musical data. O_S are objects representing the songs represented by lyrics feature, and O_A are artist objects that are formed by categorical descriptors average.

3.2 Unsupervised Graph Convolutional Networks

Graph Neural Networks offer resources to learn representations for nodes based on network topology. The representation learned is the result of a combination of neighbor representations. In summary, the representation learning problem can be denoted as a mapping function learning, $m : o_i \rightarrow z_{o_i} \in R^d$, where $m(.)$ represents the mapping function for all $o \in O$, and z_{o_i} is the newly learned representation for graph nodes existent in a space with d dimensions. In this context, initially, the objects are represented by your original features. Then, the neighboring nodes are obtained from the heterogeneous network topology, indicated by an adjacency matrix. We have learned the representation z for each object o in each GCN layer through the aggregation and combination steps. The aggregation step ($AGG(.)$) concatenates neighbors' features from a reference node, while the combination step computes an average on concatenated representation and updates the node features. Equation 1 defines the information aggregation step for node v neighbors, while 2 indicates the representation update for node v.

$$h_{N_v}^{(k)} = AGG^{(k)}(h_u^{(k-1)}, \forall u \in N(v)) \qquad (1) \qquad\qquad h_v^{(k)} = \sigma^{(k)}(h_v^{(k-1)}, h_{N_v}^{(k)}) \qquad (2)$$

in which $h(k)$ is the output in k^{th} neural network layer h, and N_v indicates the neighbors for node v.

In particular, we used an unsupervised GCN to learn the final song representation. The difference in our proposal is that the learned representation does not depend on a specific task, i.e., it can be applied in multiple downstream MIR tasks. We extend the loss function for homogeneous scenarios presented in [Hamilton et al. 2017] for the learning process to a heterogeneous graph. The representation learned for each node is based on the relations between positive and negative nodes. Given a node, the positive node are those that have a relation in the graph, and the negative nodes are randomly defined, both must have an equal type. There are no direct relations between the songs, so we exploit the node itself and the self-loop as positive. Equation 3 formulates the loss:

$$J_G(z_v) = \sum_{m}^{M} -log(\sigma(z_v^T z_u)) - Q.E_{u_n \ P_n(u)} log(\sigma(-z_v^T z_{u_n})) \qquad (3)$$

in which M is the set of node modalities, therefore $m \in \{artist, song\}$, σ is the sigmoid function, z_v and z_u demonstrate the node representations in the final output layer for node v and positive nodes u, or direct neighbors nodes, respectively. The term z_{u_n} indicates the node representation in the output layer for the negative node randomly sampled. The first term in the equation indicates that for node v, we need to be embedded closer to node u. The second term otherwise indicates that the negated dot product of negative should be maximized. $E_{u_n\ P_n(u)}$ formulates that the negative nodes are defined from a negative sampling approach, and Q defines the negative sample number. In our experiments was used one positive and negative node in the learning process.

We evaluated a large set of architecture and parameter variations for the GCN in our proposal, detailed in section 4. Then, we build inductive and transductive scenarios for GCN to explore real scenarios in music data. In transductive, we assume that all instances are already known for training the GCN, while in the inductive scenario, we separate a test (hit and non-hit songs) and a train (hit songs) set for the GNN. Finally, we generate the embeddings for all nodes after training the GNN.

3.3 One-class learning

One-Class Learning train with only one class and predict examples as belonging to the interest class or not [Emmert-Streib and Dehmer 2022]. We can define OCL as [Gôlo et al. 2021]:

$$\text{Class of } \boldsymbol{s}_i = \begin{cases} y_i \geq threshold \rightarrow \text{Hit Song} \\ y_i < threshold \rightarrow \text{Non-hit Song} \end{cases} \tag{4}$$

in which $\boldsymbol{s}_i$ is the song, and we define its class by comparing a y_i value with a threshold. y_i indicates how close the song is to belonging to the interest class, i.e., to be a hit song. In addition, to classify a song as hit or non-hit, we will indicate how hit a song will be. Suppose we used a traditional regressor in the OCL scenario. In that case, we could have limitations because the regressor predicts only hit values in the interest class hits range since the regressor will be trained only with hit songs values. Therefore, the regressor could not predict the correct popularity of a non-hit song. Thus, we choose a one-class algorithm that we can use to classify a song as a hit or not and obtain a value to represent the song's hit. In this sense, we choose the One-Class Support Vector Machines (OCSVM).

The OCSVM from [Tax and Duin 2004] learns a hypersphere to involve the interest examples. The center of the hypersphere is defined by Equation 5, in which $\boldsymbol{\mu} \in U$ is a possible center in the feature space U associated with the function kernel φ, $\varphi(\boldsymbol{s}_i)$ maps a song $\boldsymbol{s}_i$ into another feature space defined according to the kernel chosen, and $\boldsymbol{\mu}_{(c)}$ is the center of the hypersphere in which the highest distance between $\varphi(\boldsymbol{s}_i)$ to $\boldsymbol{\mu}_{(c)}$ is minimal. After defining the center, OCSVM learns the hypersphere radius r through Equation 6 subject to Equation 7, in which $\varepsilon_{\boldsymbol{s}_i}$ is the external distance between $\varphi(s_i)$ and the surface of the hypersphere, and $\nu \in (0, 1]$ defines the smoothness level of the hypersphere volume. Finally, a new example is classified as a hit song if its distance from the center is lesser than the radius, i.e., if the instance is inside the hypersphere.

$$\boldsymbol{\mu}_{(c)} = \arg\min_{\boldsymbol{\mu} \in U} \max_{1 \leq i \leq m} \|\varphi(\boldsymbol{s}_i) - \boldsymbol{\mu}\|^2 \tag{5}$$

$$\min_{\mu, \varphi, r} \quad r^2 + \frac{1}{m} \sum_{i=1}^{m} \frac{\varepsilon_{\boldsymbol{s}_i}}{\nu} \tag{6}$$

$$\|\varphi(\boldsymbol{s}_i) - \boldsymbol{\mu}_{(c)}\|^2 \leq r^2 + \varepsilon_{\boldsymbol{s}_i} \quad \forall i = 1, ..., m. \tag{7}$$

We adopted the OCSVM to predict how hit a song will be. The value predicted is directly proportional to the distance of $\boldsymbol{s}_i$ to the hypersphere, i.e., $\delta(\boldsymbol{s}_i)$. If the $\boldsymbol{s}_i$ is inside the hypersphere, we add a positive sign to the distance, and if $\boldsymbol{s}_i$ is outside the hypersphere, we add a negative sign for the distance. Thus, the higher this distance, the closer to the center is $\boldsymbol{s}_i$ (high hit song score), and the farther to the center is $\boldsymbol{s}_i$ (low hit song score), respectively. After obtaining the distances of all songs, we normalize the distances. The normalization can be defined as $\Delta(\delta(\boldsymbol{s}_i), min, max)$, in which we normalize the distance between min and max. Equation 8 defines the hit value for a song, in which the best value is 1 and the worst is 0.

$$\text{Popularity of } \boldsymbol{s}_i = \begin{cases} \delta(\boldsymbol{s}_i) \geq 0 \rightarrow \Delta(\delta(\boldsymbol{s}_i), 0.5, 1) \\ \delta(\boldsymbol{s}_i) < 0 \rightarrow \Delta(\delta(\boldsymbol{s}_i), 0, 0.5) \end{cases} \tag{8}$$

4. EXPERIMENTAL EVALUATION

We used two baselines to compare with our proposal. One is the music lyric representation obtained from the Bidirectional Encoder from Transformers model pre-trained with music lyrics (BERT)[1] [Devlin et al. 2019]. The second baseline is the concatenation of the BERT representation with the 12 features normalized by the artist (BERT+Art). We propose the song representation generated by the GNN (GNN-Song) and a GNN-Song-Art that concatenates the GNN representation for the artist and song. The initial representations for the song in the graph are the BERT representation. BERT-Music is parameter-free. The parameters of our proposal and baselines were: dimensions of the layers = $\{[64, 32], [64, 16], [128, 64, 32], [128, 64, 16], [256, 128, 64, 32], [256, 128, 64, 16], [512, 128, 64, 32]$ $, [512, 256, 128, 64, 16]\}$, learning rate = $\{0.01, 0.001, 0.0001, 0.00001\}$, optimization algorithm Adam, epochs = $\{10, 50, 100, 250, 400, 500\}$, and hyperbolic tangent activation function. Finally, the parameters of OCSVM were: kernel = $\{RBF\}$, $\nu = \{0.0001, 0.001, 0.005, 0.01, 0.05\}$ and $0.1 * \nu, \nu \in [1..9]\}$, e $\gamma = \frac{1}{n}$, in which n is the dimension of the input data.

4.1 Music Dataset and Evaluation Criteria

We use the proceeds of the 5-Fold Cross-Validation for One-Class Learning, where we divided the interest class into 5 folds, using 4 folds to train and the remaining fold and the non-interest class to test [Gôlo et al. 2022]. Then, we evaluate the representation in classification (Precision, Recall, and F_1-score metrics) and regression scenarios (mae, mse, and r^2).

We use a public dataset that concentrates on hit songs from Spotify got between 2000 and 2019. This dataset contains artists and song titles annotated with genre and the hit song score.The hit song score ranges between 0 and 89. We remove the year and explicit attributes and use twelve features extracted from the song's acoustic information to represent the artists. We represent the songs by the lyrics. To that end, we enriched this set by adding information from lyrics and the similarity relations among artists. For this process, we used the API provided by the Brazilian music portal *Vagalume*[2]. From the song title and artist, we get the lyrics. Using the artist, we get a list of related artists for our dataset. The dataset has 1617 instances after the enrichment process and the elimination of instances with no lyrics found. We define the popularity of 50 as a threshold to define a hit song. Finally, we label the songs according to popularity. Thus, we have 1353 hit songs and 264 non-hit songs.

4.2 Results and Discussion

We report in Table I the results for all music representations adopted to handle hit song prediction problems in baseline, transductive and inductive settings. In the classification scenario, we want to compute the performance to predict the hit correctly for each representation. Finally, in the regression scenario, we measure the error based on the difference between the real and prediction popularity.

We can see that the results obtained using our multimodal graph-based representation outperform the baseline representations. This result reinforces the relevance of the representation learning process for multimodal data. Furthermore, the results show that complementary latent information impacts data discrimination, but simple concatenations between features do not evidence them.

Regarding the scenarios for graph construction, the transductive scenario achieved the best results for classification and regression. This result is justified because using all instances to build the graph resulted in a more connected graph, better exploring the information aggregation process between neighbors performed by GCN. On the other hand, in the inductive scenario, we split the folds randomly and lost the relations between instances that stayed in the test and train set.

[1] `https://huggingface.co/juliensimon/autonlp-song-lyrics-18753417`.
[2] `https://api.vagalume.com.br/`.

Table I. Highest F_1-Score, Precision, Recall, mae, mse and r^2 for the representation methods Baselines (B), Inductives (I) and Transductives (T). Bold values indicate that the method obtain the best value considering all methods explored.

		Classification			Regression		
		Precision	**Recall**	**F_1-Score**	**mae**	**mse**	**r^2**
B	**BERT**	0.498±0.005	0.498±0.005	0.498±0.005	25.41±0.27	1012.20±18.9	-0.318±0.031
	BERT+Art	0.499±0.005	0.499±0.005	0.499±0.005	25.38±0.27	1010.89±19.4	-0.315±0.032
I	**GNN-Song**	0.589±0.033	0.587±0.033	0.585±0.035	24.98±0.55	980.87±31.62	-0.276±0.032
	GNN-Song-Art	0.573±0.008	0.572±0.009	0.571±0.011	24.12±0.29	954.68±27.56	-0.242±0.040
T	**GNN-Song**	0.628±0.006	**0.627±0.006**	**0.627±0.007**	22.79±0.20	891.73±10.4	-0.160±0.018
	GNN-Song-Art	**0.634±0.019**	0.626±0.023	0.619±0.031	**18.24±0.30**	**625.02±31.0**	**0.19±0.038**

Regarding the GNN-Song and GNN-Song-Art representations in the transductive scenario, for classification, aggregating the artist information in the music representation through relationships is more effective than concatenating the representation of both after the learning process. However, for regression, we need more information to estimate musical popularity. These results allow us to infer that our proposal incorporates the semantic information between the features learning a more discriminative representation, but a larger volume of related data is needed to build a more connected graph.

5. CONCLUSIONS

This work presents an approach to learning graph-based multimodal representation for music data applied to the hit-song prediction problem. We propose using an unsupervised GCN with an extended loss function for a data scenario with multimodal and heterogeneous features. Our approach's differential is using information from the network relationships to learn a representation for the nodes without dependence on labels. The hit song prediction problem was formulated as a one-class learning problem, reducing the dependence on annotated data for the non-interest class. The results evidenced the graph-based representations' ability to incorporate semantic information from multiple features and outperform unimodal representations or generated from feature concatenations.

Our work was limited to a strategy of selecting positive and negative instances in addition to being evaluated on a single dataset, which does not measure the generalization power of the approach. So, in future work, we will expand on the strategies for selecting positive and negative instances in the loss function in other datasets. Finally, we hope to incorporate other acoustic features into the representation to add more semantic information to the learning process and learn a more robust musical representation. The source code and used dataset are available in the public repository[3].

REFERENCES

Ali, I. and Melton, A. Graph-based semantic learning, representation and growth from text: A systematic review. In *2019 IEEE 13th ICSC*. IEEE, Newport Beach, CA, USA, pp. 118–123, 2019.

Bengio, Y., Courville, A., and Vincent, P. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence* 35 (8): 1798–1828, 2013.

Bertoni, A. A. et al. *Avaliação de características e previsão de sucesso de canções populares brasileiras por meio de aprendizado de máquina*. M.S. thesis, Universidade Federal de Goiás, 2021.

Chen, W., Keast, J., Moody, J., Moriarty, C., Villalobos, F., Winter, V., Zhang, X., Lyu, X., Freeman, E., Wang, J., Kai, S., and Kinnaird, K. M. Data usage in mir: History & future recommendations. In *International Society for Music Information Retrieval Conference*. ISMIR, Delft, The Netherlands, pp. 25–32, 2019.

da Silva, A. C. M., do Carmo, P. R. V., Marcacini, R. M., and Silva, D. F. Instance selection for music genre classification using heterogeneous networks. *Simpósio Brasileiro de Computação Musical* vol. 18, pp. 11–18, 2021.

da Silva, A. C. M., Silva, D. F., and Marcacini, R. M. Multimodal representation learning over heterogeneous networks for tag-based music retrieval. *Expert Systems with Applications* vol. 207, pp. 1–9, 2022.

[3] `https://github.com/AngeloMendes/Unsupervised-Heterogeneous-Graph-Neural-Network-for-Hit-Song-Prediction-through-One-Class-Learning`

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In *2019 NAACL*. ACL, Minnesota, pp. 4171–4186, 2019.

Emmert-Streib, F. and Dehmer, M. Taxonomy of machine learning paradigms: A data-centric perspective. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* vol. e1470, pp. 25, 2022.

Ge, Y., Wu, J., and Sun, Y. Popularity prediction of music based on factor extraction and model blending. In *2020 2nd ICEMME*. IEEE, Chongqing, China, pp. 1062–1065, 2020.

Gôlo, M., Caravanti, M., Rossi, R., Rezende, S., Nogueira, B., and Marcacini, R. Learning textual representations from multiple modalities to detect fake news through one-class learning. In *Proc. of the Brazilian Symposium on Multimedia and the Web*. ACM, Belo Horizonte, Brazil, pp. 197–204, 2021.

Gôlo, M. P., Araújo, A. F., Rossi, R. G., and Marcacini, R. M. Detecting relevant app reviews for software evolution and maintenance through multimodal one-class learning. *Inf. Software Technology* vol. 151, pp. 1–12, 2022.

Hamilton, W., Ying, Z., and Leskovec, J. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). Vol. 30. Curran Associates, Inc., LONG BEACH, CA, USA, pp. 1–11, 2017.

Herremans, D., Martens, D., and Sörensen, K. Dance hit song prediction. *New Music Res.* 43 (3): 291–302, 2014.

Hiller, R. S. and Walter, J. M. The rise of streaming music and implications for music production. *Review of Network Economics* 16 (4): 351–385, 2017.

IFPI. Ifpi issues global music report 2022. `https://globalmusicreport.ifpi.org/`, 2022. Accessed: 09-12-2022.

Karydis, I., Gkiokas, A., Katsouros, V., and Iliadis, L. Musical track popularity mining dataset: Extension & experimentation. *Neurocomputing* vol. 280, pp. 76–85, 2018.

Kim, J., Urbano, J., Liem, C., and Hanjalic, A. One deep music representation to rule them all? a comparative analysis of different representation learning strategies. *Neural Computing and Applications* 32 (4): 1067–1093, 2020.

Knees, P. and Schedl, M. A survey of music similarity and recommendation from music context data. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 10 (1): 1–21, 2013.

Korzeniowski, F., Oramas, S., and Gouyon, F. Artist similarity with graph neural networks. In *Proc. of International Society for Music Information Retrieval*. ISMIR, Online, 2021.

Krohn-Grimberghe, L. How streaming technology impacts music production and consumption. In *Classical Concert Studies*. Routledge, New York, EUA, 30, pp. 296–308, 2021.

Martín-Gutiérrez, D., Peñaloza, G. H., Belmonte-Hernández, A., and García, F. Á. A multimodal end-to-end deep learning architecture for music popularity prediction. *IEEE Access* vol. 8, pp. 39361–39374, 2020.

Melo, D. d. F. P., Fadigas, I. d. S., and Pereira, H. B. d. B. Graph-based feature extraction: A new proposal to study the classification of music signals outside the time-frequency domain. *PLOS ONE* 15 (11): 1–26, 11, 2020.

Pareek, P., Shankar, P., Pathak, M. P., and Sakariya, M. N. Predicting music popularity using machine learning algorithm and music metrics available in spotify. *Center for Development Economic Studies* 9 (11): 10–19, 2022.

Sawant, S. S. and Prabukumar, M. A review on graph-based semi-supervised learning methods for hyperspectral image classification. *The Egyptian Journal of Remote Sensing and Space Science* 23 (2): 243–248, 2020.

Shi, C. Heterogeneous graph neural networks. In *Graph Neural Networks: Foundations, Frontiers, and Applications*, L. Wu, P. Cui, J. Pei, and L. Zhao (Eds.). Springer Singapore, Singapore, 16, pp. 351–370, 2022.

Silva, A. C. M., Silva, D. F., and Marcacini, R. M. Heterogeneous graph neural network for music emotion recognition. In *Proc. of International Society for Music Information Retrieval*. ISMIR, Bengaluru, India, 2022.

Simonetta, F., Ntalampiras, S., and Avanzini, F. Multimodal music information processing and retrieval: Survey and future challenges. In *2019 International Workshop on Multilayer Music Representation and Processing (MMRP)*. IEEE, Milan, Italy, pp. 10–18, 2019.

Singhi, A. *Lyrics matter: Using lyrics to solve music information retrieval tasks*. M.S. thesis, Uni. of Waterloo, 2015.

Singhi, A. and Brown, D. G. Can song lyrics predict hits. In *Proceedings of the 11th International Symposium on Computer Music Multidisciplinary Research*. University of Waterloo, Plymouth, UK, pp. 457–471, 2015.

Song, Y., Dixon, S., and Pearce, M. A survey of music recommendation systems and future perspectives. In *9th international symposium on computer music modeling and retrieval*. Vol. 4. Springer, London, pp. 395–410, 2012.

Tax, D. and Duin, R. Support vector data description. *Machine Learning* 54 (1): 45–66, 2004.

Tax, D. M. J. *One-class classification: Concept learning in the absence of counter-examples*. Ph.D. thesis, Technische Universiteit Delft, 2001.

Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., and Yu, P. S. A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems* 32 (1): 4–24, 2021.

Xia, F., Sun, K., Yu, S., Aziz, A., Wan, L., Pan, S., and Liu, H. Graph learning: A survey. *IEEE Transactions on Artificial Intelligence* 2 (02): 109–127, apr, 2021.

Yang, C., Xiao, Y., Zhang, Y., Sun, Y., and Han, J. Heterogeneous network representation learning: A unified framework with survey and benchmark. *Transactions on Knowledge and Data Engineering* vol. PP, pp. 1–1, 2020.

Zangerle, E., Vötter, M., Huber, R., and Yang, Y.-H. Hit song prediction: Leveraging low-and high-level audio features. In *ISMIR*. ISMIR, Delft, Netherlands, pp. 319–326, 2019.