ISSN 1870-4069

NLP Applied to Musical Harmonic Structures for Music Emotion Recognition

Leonardo Daniel Villanueva Medina, Efrén Gorrostieta Hurtado

Universidad Autónoma de Queretaro,
Mexico

lvillanueva@uaq.mx

Abstract. Music emotion recognition (M.E.R.) is a multidisciplinary field that integrates computer science, affective computing, and neuroscience elements to analyze musical features to detect emotions. Most research in this field has focused on low and mid-level features, often ignoring theoretical and harmonic aspects of music. In this work, we propose using regression-based machine learning models applied to word embeddings in harmonic structures (chords). The results indicate an RMSE of 0.0252 and an R^2 score of 0.9751 for the valence dimension, in comparison with the arousal, an RMSE of 0.1319 and an R^2 score of 0.4676. These findings indicate that incorporating theoretical and harmonic concepts enhances the performance of M.E.R models, particularly in the valence dimension, reflecting improved detection of the positivity of emotions.

Keywords: mer, word embeddings, machine learning, musical features.

1 Introduction

Music has remarkably impacted social, cultural, and political aspects. For this reason, it has been the target of many studies, one of them being the relationship between emotions and music [13] since music is a means of expression capable of evoking emotions [6].

Music Emotion Recognition (M.E.R.) has incorporated knowledge from several fields, such as computer science, affective computing, and neuroscience. It aims to analyze musical features extracted from audio signals (low and mid-level) and abstract features such as song lyrics (high-level) [13, 9, 15, 7].

Within M.E.R.'s works, two approaches for linking emotions and songs predominate. The first one attaches a general emotion to the whole work (song-level), a static approach. The second, dynamic approach, focuses on detecting the music emotion variations that occur through the song, namely MEDV (music emotion variation) [9, 6].

Emotional perception is complex because it involves multiple variables, such as the song or external information, such as the listener's social, cultural, and emotional context [17, 7].

Selecting the appropriate taxonomy is crucial for clearly delineating the problem as either a multi-class classification or a regression task [9]. In this

regard, there are two main approaches: categorical taxonomies, which represent emotions through adjectives (such as Hevner's model [11]), and dimensional taxonomies, which represent emotions employing numerical values (such as Russell and Thayer's models [18, 16, 20]). The dimensional representation is organized based on two emotional axes: valence and arousal. Valence represents how pleasant an emotion is, and arousal represents excitement [7].

Traditional M.E.R. works are commonly based on the analysis of low-mid-level features. Therefore, as several works have proposed, information and features directly linked to emotions are needed [15, 17, 5, 22, 8].

Important musical concepts, such as theory and harmony, must be understood in the music-emotion relationship [11, 19, 15]. In this context, there are two essential elements: scales and chords. A scale is a succession of notes that follow a pattern. At the same time, a chord is a sequence of more than two notes. Each note forms a chord within a scale, which serves a function, such as determining the scale mode (major or minor) or indicating a transition or end/rest of a segment [10].

Similarly, according to Steinbeis' experiments [19], the listener expects a sense of closure by resting chords at the end of a harmonic progression. Replacing these with transitional chords can alter that emotional experience and change the expectation of the work's end.

Additionally, a certain similarity between the representation of chords and natural language has been pointed out, thus enabling the application of Natural Language Processing (NLP) techniques [12, 8].

This work introduces a technique combining harmony ideas and musical theory to improve music emotion recognition. We present adapted word embeddings to song chords based on natural language processing (NLP) methods. This method chooses suitable machine learning models to interpret the chord embeddings, facilitating a simpler, detailed analysis. Finally, the main goal of this work is to predict valence and arousal values to ensure that we can precisely detect the emotions perceived in music.

2 Background

The traditional methodology of M.E.R works is based on analyzing low-mid-level features through machine learning models. Nevertheless, multimodal strategies that assemble deep learning, NLP, and traditional techniques have been adopted, generating robust models.

In this regard, Panda et al.[15] underline the need for design elements that capture the music-emotion relation. This work is based on a novel set of features and employs a support vector machine model for the multi-classification problem, reaching a 76.4% value for an f1-score metric.

On the other hand, Yang's work [21] carries out emotion recognition through a Back Propagation algorithm (BP). The model's input was a set of six different low-level features. Enhance the BP algorithm with metaheuristic techniques, specifically an artificial bee colony algorithm. The results indicate an MAE of

0.8872, RMSE of 0.1066, and an R^2 of 0.4606 for the valence dimension. For the arousal dimension, an MAE of 0.9156, RMSE of 0.1322, and R^2 of 0.6687.

Some multi-modal approaches merge low-mid-level features with high-level information through deep learning and NLP models. In [17], the work addressed emotion recognition by analyzing two types of features. Researchers used CNN models for the low-level features (in spectrogram form) and applied several NLP methods to the song lyrics, reaching the best results with BERT embeddings.

On the other hand, in work [3], the focus is on the MER task using source separation from the PMEmo dataset, where the audio is split into four tracks (vocals, bass, drums, and other), from which spectrograms were extracted.

At the same time, efforts have been made to detect emotions using chords and harmony progressions. Cho's work [5] performs emotion detection based on MIDI and audio files. This work uses a chord matrix (coding the chord position) in combination with low-level features to predict the valence and arousal values through an SVR model. This results in an MSE of 0.67 for the valence dimension with the MIDI files and an MSE of 0.65 for the arousal using the audio files. In contrast, Zhang's work [22] tackles the M.E.R problem through statistical methods. A database that bonds emotion to a set of chords and identifies the chords by the analysis of low-level features.

NLP methods are not limited to song lyrics analysis. To demonstrate that embeddings can describe chords like music theory does, Lahnala's work [12], for example, uses Word2Vec predictive embeddings to capture the association between chords. On the other hand, Greer [8] improves the precision of music emotion classification by approaching the problem as a multi-classification problem by combining lyrics and chords into shared vectors.

3 Methods

Figure 1 illustrates the overall methodology diagram used in this work. The following sections provide a detailed explanation of each step.

3.1 Dataset

This study uses two well-known datasets in the M.E.R field: PMEMO [23] and MEDV [1]. Both datasets include static dimensional annotations composed of valence and arousal axes whose values are normalized from 0 to 1. The MEDV dataset includes 1802 MP3 audio files for 45 seconds. The PMEMO dataset has 767 audio MP3 files whose duration may vary. The sample frequency is the same in both datasets, 44.1 kHz.

3.2 Audio File Conversion

This work employs the Python libraries Librosa [14], Soundfile, and FFmpeg, which have functions that boost reading, conversion between different types of audio files, and writing of new files. The audio file format was converted from MP3 to new WAV files.

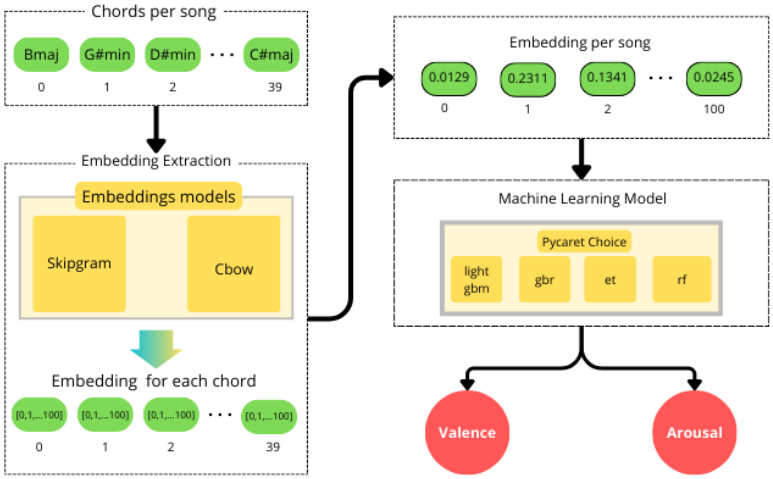

Fig. 1. Methodology implemented for the music emotion recognition (dimensional taxonomy).

3.3 Chord Detection and Data Augmentation

Chord detection was performed using the MADMOM [4] library (MADMOM only works with WAV files). In this way, chords were retrieved from the 2569 audio files. However, MADMOM only recognizes major and minor chords.

Chord transposition was used for data augmentation, a musical technique that increases or decreases the notes of a chord. Thus, transposition was applied in half-tone, whole-tone intervals (ascending and descending). The table 1 shows an example of this technique. With the increased data, the corpus was extended to 12845 progressions. Even with the limitation of MADMOM and the fact that there are only 12 sounds in Western music, the final dictionary is restricted to 24 chords.

3.4 Embeddings

Cooccurrence-based embeddings, such as word2vec, capture words' semantic and syntactic information and represent it in a vector space. In this way, each word forms a vector $\mathbb{R}^N$ where the dimension $N \geq 100$. Thus, based on their proximity in the plane, it can be known which words share context [17]. Skip-gram and CBOW are methods of such embeddings. Skip-gram predicts the context using the core word, while CBOW identifies the core word from the surrounding context [12, 8].

Based on Lanhala's work [12], CBOW and skip-gram were used as embeddings. The size of the embeddings was set to 100 and 200, while the

Table 1. Data augmentation for the song "I Have Questions" by "Camila Cabello" from the PMEMO dataset (just the first four chords).

Chords	Arousal	Valence	Type
G#min Emaj F#maj G#min	0.7375	0.7375	whole-tone Down
Amin Fmaj Gmaj Amin	0.7375	0.7375	half-tone Down
A#min F#maj G#maj A#min	**0.7375**	**0.7375**	**Source**
Bmin Gmaj Amaj Bmin	0.7375	0.7375	half-tone Up
Cmin G#maj A#maj Cmin	0.7375	0.7375	whole-tone Up

windows were 5, 10, and 20. Each song was set to a maximum length of 40 chords, truncating with zeros if necessary, and the embeddings for each chord were averaged to obtain a unique embedding.

Subsequently, a PCA dimensionality reduction algorithm was applied to visualize the relationship between chords in a two-dimensional plane. Figure 2 shows a circular arrangement similar to the "circle of fifths." This structure indicates the relationship and similarity between chords [12, 10]. Despite the limited dictionary, the representation of the relationship between major and minor chords is inadequate in the plan.

3.5 Machine Learning Model

A machine learning model has been implemented for emotion recognition. PyCaret library [2] was employed to automate the model selection and validation process for a regression task. The experimental setup used a train size of 80%, with the remaining 20% reserved for validation. The split was performed randomly to ensure a representative data distribution across both subsets. Various regression algorithms were trained and evaluated during the model comparison phase. Model performance was assessed based on the mean absolute error (MAE), root mean squared error (RMSE), and the coefficient of determination (R^2). Models were ranked according to these metrics, and the best-performing model was selected based on the primary evaluation criteria. Finally, a 10-fold cross-validation was performed to validate the best model chosen for valence and arousal.

The models that consistently performed best were the **Gradient Boosting Regressor (GBR)**, **Light Gradient Boosting Machine (LIGHTBGM)**, **Random Forest (RF)**, and **Extra Trees Classifier (ET)**. The model takes the unique embeddings as input and produces the valence and arousal values as output.

The RMSE, MAE, and R^2 metrics are the most commonly used in regression problems for emotion recognition [21].

The root mean square error (RMSE) measures the dispersion of errors and penalizes extreme values (equation 1, [21]). The mean absolute error (MAE) calculates the average difference between the actual and predicted values (equation 2, [21]). The coefficient of determination R^2 evaluates how well the

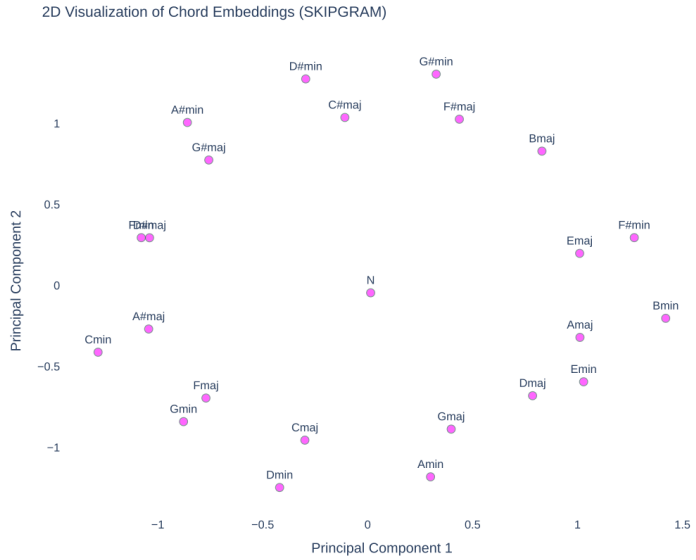

Fig. 2. Vector representation of the relationships captured by the single chord embeddings. The minor chords form the outer circle, and the major chords form the inner circle.

predicted values match the actual values; an R^2 close to 1 indicates better accuracy (equation 3, [21]).

Where y_i are the observed values, $\hat{y}_i$ is the model predictions, and $\bar{y}$ represents the mean of the observed values:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}, \tag{1}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right|, \tag{2}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}. \tag{3}$$

4 Results

4.1 Embeddings

The Skip-gram and CBOW models retrieved the embedding of each dictionary element. Both methods capture the relationship between chords just as music theory does. However, when evaluating the cosine similarity of the five most similar chords (figure 3), CBOW showed lower values, so Skip-gram embeddings were used exclusively in the subsequent experiments.

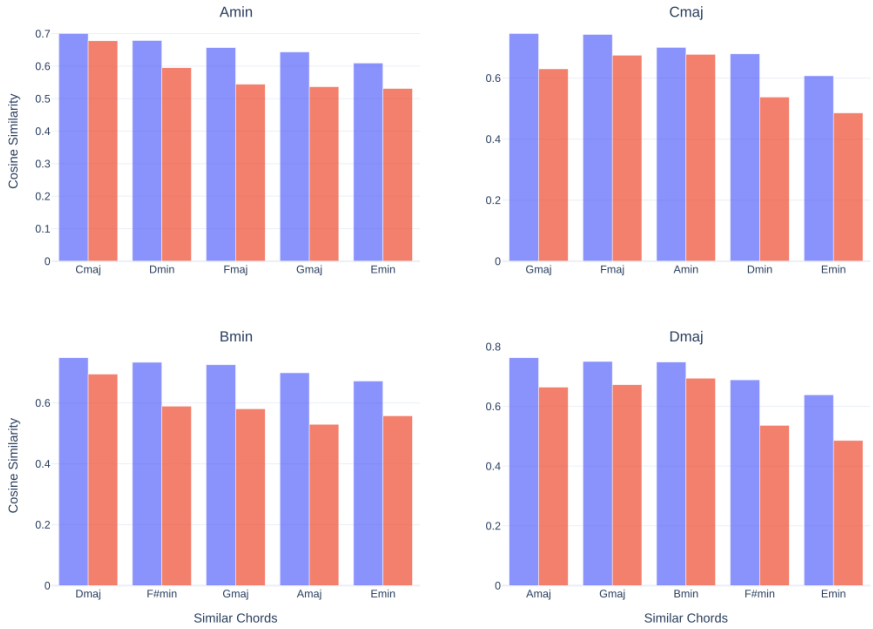

Fig. 3. Blue: skip-gram. Red: Cbow. It shows the five most similar chords: A minor, C major, B minor, and D major.

4.2 Emotion Recognition

The search for the best machine learning models with PyCaret yields metrics from Tables **??** and 2, which compare the models grouped by valence and arousal.

In the arousal dimension, the GBR and LIGHTGBM models demonstrate comparable results. The difference between errors is slight and does not exceed 0.0006 points for the MAE and 0.0019 for the RMSE. The models differ in the R^2, with the highest value, 0.4620, and the lowest, 0.4573.

According to Table 2, in the dimension of arousal, the error is comparatively reduced when working with embeddings of size 100, and the value of R^2 is higher than with embeddings of size 200. The best result is obtained with embeddings of size 100 and a window of 10, employing the lightgbm model.

In the valence dimension, the LIGHTGMB model achieves fewer errors and better scores on the R^2. In general, for this axis, better results are obtained with embeddings of size 200. However, the best result is achieved with 100 embeddings and a 5-window, with an MAE of 0.0175, an RMSE of 0.0260, and an R^2 of 0.9733.

Table **??** shows how, in the embedding vectors with larger size (200), better results are achieved with larger contextual windows (10 and 20). On the contrary, in embeddings with reduced size (100), the best result is achieved with a small contextual window (5).

Table 2. Comparison of metrics for Arousal with different models and configurations (Dim: 100 and Dim: 200).

	Dim: 100			**Dim: 200**		
	MAE	**RMSE**	**R^2**	**MAE**	**RMSE**	**R^2**
Win: 5						
gbr	**0.1062**	**0.1330**	**0.4585**	0.1064	0.1332	0.4571
lightgbm	0.1062	0.1332	0.4566	**0.1061**	**0.1331**	**0.4573**
rf	0.1062	0.1339	0.4513	0.1067	0.1343	0.4475
Win: 10						
gbr	0.1062	0.1330	0.4581	**0.1063**	**0.1331**	**0.4573**
lightgbm	**0.1058**	**0.1326**	**0.4620**	0.1061	0.1332	0.4571
rf	0.1064	0.1341	0.4493	0.1065	0.1339	0.4506
Win: 20						
gbr	0.1064	0.1332	0.4564	**0.1062**	**0.1331**	**0.4574**
lightgbm	**0.1061**	**0.1330**	**0.458**2	0.1065	0.1337	0.4529
rf	0.1064	0.1349	0.4508	0.1065	0.1342	0.4484

The values of MAE, RMSE, and R^2 for Arousal. 'Dim' corresponds to the embedding dimension, and 'Win' corresponds to the size of the contextual window.

4.3 Cross Validation

The models with the best results in embeddings of size 100 and window size five were selected since this configuration optimized the performance in valence. Thus, LIGHTGBM was chosen for valence and GBR for arousal. Figure 4 shows that the valence predictions are close to the actual values, while the arousal predictions are concentrated in a middle range, moving away from the baseline. Each model was retrained and validated using a 10-fold cross-validation strategy.

The results obtained from this experiment are summarized in Tables 4 and 5. For the arousal dimension, the model achieved a mean MAE of 0.1049, a mean RMSE of 0.1319, and a mean R^2 of 0.4676. In the case of valence, the model achieved a mean MAE of 0.0167, a mean RMSE of 0.0252, and a mean R^2 of 0.9751, with relatively low standard deviations across all folds.

A subset of 'simple' songs (those with reduced progressions and only major and minor chords) was evaluated as a comparison. Figure 5 shows how the predictions follow the previous trends in both dimensions, highlighting a better performance in arousal for these songs.

4.4 Comparison

Table 6 compares the best results of selected state-of-the-art works with the best model proposed in this study. The results highlight that the analysis of harmonic structures achieves better performance than the analysis of purely acoustic

Table 3. Comparison of metrics for Valence with different models and configurations (Dim: 100 and Dim: 200).

	Dim: 100			**Dim: 200**		
	MAE	**RMSE**	**R^2**	**MAE**	**RMSE**	**R^2**
Win: 5						
lightgbm	**0.0175**	**0.0260**	**0.9733**	**0.0332**	**0.0461**	**0.9160**
et	0.0186	0.0277	0.9696	0.0349	0.0486	0.9069
rf	0.0188	0.0280	0.9691	0.0351	0.0488	0.9060
Win: 10						
lightgbm	**0.0328**	**0.0457**	**0.917**6	**0.0181**	**0.0272**	**0.9708**
et	0.0343	0.0478	0.9098	0.0193	0.0289	0.9671
rf	0.0345	0.0483	0.9098	0.0196	0.0294	0.9660
Win: 20						
lightgbm	**0.0327**	**0.0463**	**0.9192**	**0.0175**	**0.0268**	**0.9717**
et	0.0342	0.0473	0.9118	0.0184	0.0283	0.9684
rf	0.0345	0.0478	0.9099	0.0186	0.0286	0.9678

The values of MAE, RMSE, and R^2 for Valence. 'Dim' corresponds to the embedding dimension, and 'Win' corresponds to the size of the contextual window.

Table 4. Cross-validation results for the arousal dimension after data augmentation.

Fold	**MAE**	**RMSE**	**R^2**
0	0.1073	0.1343	0.4541
1	0.1024	0.1276	0.4609
2	0.1023	0.1284	0.4850
3	0.1071	0.1353	0.4693
4	0.1014	0.1282	0.4961
5	0.1025	0.1291	0.4956
6	0.1052	0.1314	0.4574
7	0.1099	0.1384	0.4433
8	0.1056	0.1335	0.4609
9	0.1054	0.1331	0.4534
Mean	**0.1049**	**0.1319**	**0.4676**
Std	**0.0026**	**0.0034**	**0.0175**

The table reports the MAE, RMSE, and R^2 metrics obtained from 10-fold cross-validation for the gbr model for the arousal dimension.

characteristics, at least for recognizing the valence (positivity or negativity) a listener perceives.

Skip-gram predictive embeddings facilitate pattern learning in artificial intelligence models, enhancing the analysis of harmonic structures for emotion recognition. Thus, combining high-level harmonic features informed by music

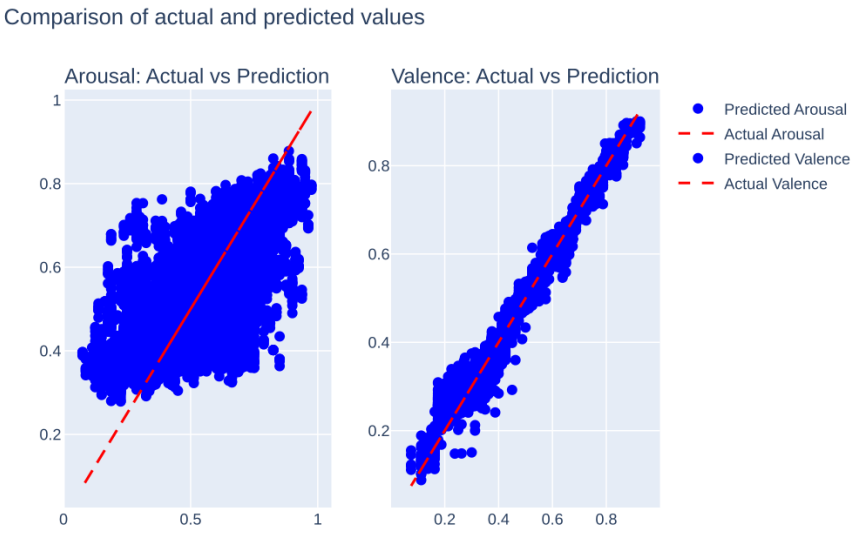

Fig. 4. Behavior of the model in the dimensions of arousal and valence.

Table 5. Cross-validation results for the valence dimension after data augmentation.

Fold	MAE	RMSE	R^2
0	0.0166	0.0255	0.9733
1	0.0166	0.0253	0.9769
2	0.0163	0.0233	0.9786
3	0.0162	0.0239	0.9760
4	0.0165	0.0253	0.9756
5	0.0169	0.0259	0.9740
6	0.0168	0.0251	0.9751
7	0.0168	0.0243	0.9769
8	0.0173	0.0273	0.9710
9	0.0166	0.0257	0.9733
Mean	**0.0167**	**0.0252**	**0.9751**
Std	**0.0003**	**0.0011**	**0.0021**

The table reports the MAE, RMSE, and R^2 metrics obtained from 10-fold cross-validation for the lightgbm model for the valence dimension.

theory with natural language processing techniques proves more efficient than traditional MER approaches for recognizing the valence dimension.

However, low-level acoustic features remain superior for recognizing the arousal dimension. Acoustic-based systems typically consider a wide range of features such as pitch, timbre, and rhythm. In contrast, the harmonic analysis

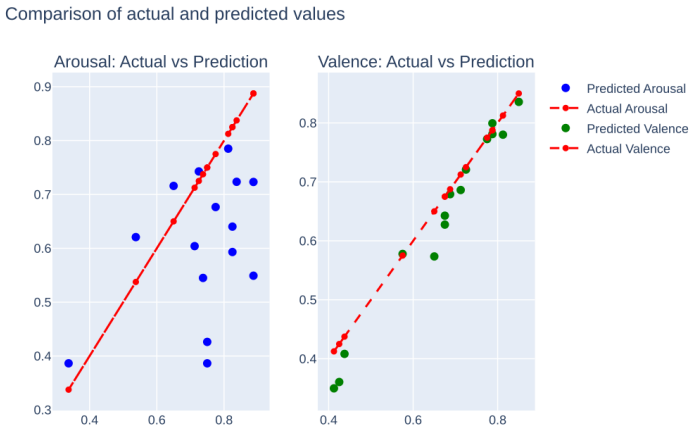

Fig. 5. Comparison between actual and predicted values for the easy songs set.

in this work was limited to only two chord modes (major and minor), focusing on basic chord structures.

Table 6. Comparison of state-of-the-art results and the present work. RMSE and R^2 metrics for Arousal and Valence are shown.

	RMSE		**R^2**		**T.C**
	Arousal	**Valence**	**Arousal**	**Valence**	
Y-H. Cho [5]	0.806	1.104	—	—	Chords
PMEMO [23]	**0.102**	0.124	—	—	Acoustics
EmoMucs [3]	0.2307	0.2373	0.6046	0.4584	Acoustics
Yang [21]	0.1322	0.1066	0.6687	0.4606	Acoustics
Proposed	0.1319	**0.0252**	0.4676	0.9751	Chords

Note: The table shows the best results of each work. Column T.C references the main feature that drives the models. Works marked with --- did not report the R^2 metric.

5 Discussion and Conclusion

The predictive skip-gram embeddings could capture the music-theoretic relationship between major and minor chords, similar to Lahnala's work [12]. When representing the embeddings in the plane, the distribution of the embeddings is almost similar to the circle of fifths of music theory, failing in the distribution of the outer circle (relative minors). However, the data set used

in this work was limited since, unlike Lahnala's or Greer's work [12, 8], only major and minor chords were extracted, resulting in a reduced chord dictionary. Even so, this only affects the representation in the plane since when calculating the cosine similarity (figure 3), we observe how the embeddings manage to capture the relationship between relative major and minor. Thus, skip-gram boosted with data augmentation can capture relationships between chords as music theory dictates.

Finally, it can be observed how analyzing the harmonic structure of a song improves the results obtained by M.E.R systems based on the analysis of low-level acoustic features [23, 21], demonstrating the strong link of harmonic structure in the perception of positivity of emotions in music, furthermore, using only major and minor chords does not affect emotion recognition and boosts the results in the valence dimension because major and minor modes are often associated with feelings such as happiness and sad [11]. The model performs better than the average in the arousal dimension, as indicated by the R^2. However, the result is low compared to works such as [23, 21]. This may be so because the harmonic structure of a song does not reflect other characteristics such as rhythm, color, or degree of energy, thus determining that, in predicting the degree of intensity of an emotion, it is better to work with low-level characteristics. Nonetheless, there is significant potential for improvement, particularly by incorporating more complex chords. Additionally, implementing contextual embeddings could enhance the results further. However, this would necessitate a larger dataset.

In conclusion, analyzing harmonic structures with basic chords improves emotion recognition based on dimensional taxonomies and significantly impacts predicting valence. Nonetheless, the exclusive analysis of harmonic structures leaves aside other factors related to the arousal axis, such as rhythm, which affects the performance of machine learning models in predicting values of this axis.

As future work, we aim to explore multimodal approaches that combine chord embeddings with acoustic features derived from spectrogram representations, such as chromagrams, constant-Q transforms (CQT), and mel-spectrograms. While chord embeddings effectively capture the harmonic structure relevant to valence prediction, spectrogram-based features offer a richer representation of the temporal and dynamic aspects of music. By integrating these modalities, we expect to enhance the model's ability to capture the energy and intensity variations associated with arousal, which are not fully reflected in harmonic content alone.

References

1. Alajanki, A., Yang, Y.-H., Soleymani, M.: Benchmarking music emotion recognition systems. PLOS ONE, (2016)
2. Ali, M.: PyCaret: An open source, low-code machine learning library in Python (April 2020), https://www.pycaret.org, pyCaret version 1.0

3. de Berardinis, J., Cangelosi, A., Coutinho, E.: The multiple voices of musical emotions: source separation for improving music emotion recognition models and their interpretability. International Society for Music Information Retrieval Conference, (2020)
4. Böck, S., Korzeniowski, F., Schlüter, J., Krebs, F., Widmer, G.: madmom: a new Python Audio and Music Signal Processing Library. In: Proceedings of the 24th ACM International Conference on Multimedia. pp. 1174–1178. Amsterdam, The Netherlands (10 2016) doi: 10.1145/2964284.2973795
5. Cho, Y.-H., Lim, H., Kim, D.-W., Lee, I.-K.: Music emotion recognition using chord progressions. In: 2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC). pp. 002588–002593. IEEE (10 2016) doi: 10.1109/SMC.2016.7844628
6. Cui, X., Wu, Y., Wu, J., You, Z., Xiahou, J., Ouyang, M.: A review: Music-emotion recognition and analysis based on eeg signals. Frontiers in Neuroinformatics, vol. 16, pp. 997282 (10 2022) doi: 10.3389/FNINF.2022.997282/BIBTEX
7. Gomez-Canon, J. S., Cano, E., Eerola, T., Herrera, P., Hu, X., Yang, Y.-H., Gomez, E.: Music emotion recognition: Toward new, robust standards in personalized and context-sensitive applications. IEEE Signal Processing Magazine, vol. 38, pp. 106–114 (11 2021) doi: 10.1109/MSP.2021.3106232
8. Greer, T., Singla, K., Ma, B., Narayanan, S.: Learning shared vector representations of lyrics and chords in music. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 3951–3955. IEEE (2019)
9. Han, D., Kong, Y., Han, J., Wang, G.: A survey of music emotion recognition. Frontiers of Computer Science, vol. 16, pp. 166335 (12 2022) doi: 10.1007/s11704-021-0569-4
10. Herrera, E.: Teoría musical y armonía moderna Vol. 2, vol. 2. Antoni Bosch editor (2022)
11. Hevner, K.: Experimental studies of the elements of expression in music. The American Journal of Psychology, vol. 48, pp. 246 (4 1936) doi: 10.2307/1415746
12. Lahnala, A., Kambhatla, G., Peng, J., Whitehead, M., Minnehan, G., Guldan, E., Kummerfeld, J. K., Çamcı, A., Mihalcea, R.: Chord embeddings: Analyzing what they capture and their role for next chord prediction and artist attribute prediction. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 12693 LNCS, pp. 171–186 (2021) doi: 10.1007/978-3-030-72914-1_12
13. Lucia-Mulas, M. J., Revuelta-Sanz, P., Ruiz-Mezcua, B., Gonzalez-Carrasco, I.: Automatic music emotion classification model for movie soundtrack subtitling based on neuroscientific premises. Applied Intelligence, vol. 53, pp. 27096–27109 (11 2023) doi: https://doi.org/10.1007/s10489-023-04967-w
14. McFee, B., Raffel, C., Liang, D., Ellis, D. P., McVicar, M., Battenberg, E., Nieto, O.: librosa: Audio and music signal analysis in python. SciPy, vol. 2015, pp. 18–24 (2015)
15. Panda, R., Malheiro, R., Paiva, R. P.: Novel audio features for music emotion recognition. IEEE Transactions on Affective Computing, vol. 11, pp. 614–626 (10 2020) doi: 10.1109/TAFFC.2018.2820691
16. Posner, J., Russell, J. A., Peterson, B. S.: The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. Development and Psychopathology, vol. 17, pp. 715–734 (7 2005) doi: 10.1017/S0954579405050340

17. Pyrovolakis, K., Tzouveli, P., Stamou, G.: Multi-modal song mood detection with deep learning. Sensors 2022, Vol. 22, Page 1065, vol. 22, pp. 1065 (1 2022) doi: 10.3390/S22031065
18. Russell, J. A.: A circumplex model of affect. Journal of Personality and Social Psychology, vol. 39, pp. 1161–1178 (12 1980) doi: 10.1037/H0077714
19. Steinbeis, N., Koelsch, S., Sloboda, J. A.: The role of harmonic expectancy violations in musical emotions: Evidence from subjective, physiological, and neural responses. Journal of Cognitive Neuroscience, vol. 18, pp. 1380–1393 (8 2006) doi: 10.1162/jocn.2006.18.8.1380
20. Thayer, R. E.: The Biopsychology of Mood and Arousal. Oxford University PressNew York, NY (9 1990)
21. Yang, J.: A novel music emotion recognition model using neural network technology. Frontiers in Psychology, vol. 12, pp. 760060 (9 2021) doi: 10.3389/FPSYG.2021.760060
22. Zhang, F., Meng, H., Li, M., Cui, R., Liu, C.: Music emotion recognition based on chord identification. In: The International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery. pp. 956–963. Springer (2020)
23. Zhang, K., Zhang, H., Li, S., Yang, C., Sun, L.: The pmemo dataset for music emotion recognition. In: Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval. pp. 135–142. ICMR '18, ACM, New York, NY, USA (2018) doi: 10.1145/3206025.3206037