CENTRO DE INVESTIGACIÓN EN MATEMÁTICAS

Maestría en Modelación y Optimización de Procesos
Simulación de Procesos Discretos (Deep Learning)
12 de diciembre de 2023

CIMAT

Análisis de emoción y sentimiento en música desde una perspectiva multimodal

Salvador Sánchez Velázquez
Lizbeth Alejandra Torres Nájera

Resumen

Este artículo se enfoca en el análisis de emoción y sentimiento en la música desde una perspectiva multimodal bajo la extracción de características. Se aborda la detección de emociones mediante las medidas de *arousal* y *valence* en obras musicales utilizando modelos que combinan información acústica y textual. Se incluyen modelos de Multimodal-Toolkit que emplean el transformer DistilBERT para situaciones con texto y datos tabulares, que en este caso, se hace uso de la base de datos de Deezer Mood Detection Dataset (DMDD), que contiene características de audio de canciones obtenidas a través del API de Spotify. Finalmente se discute la comparación del rendimiento de modelos unimodales (letra o musica) con el modelo multimodal que combina ambos enfoques.

I. Introducción

Una tarea muy popular en el área de recuperación de información musical (MIR) es la detección de emoción (mood) y sentimiento en obras musicales. En este proyecto se aborda esta tarea bajo una perspectiva multimodal, explorando modelos que incluyen representaciones de información acústica y de texto para la identificación de emociones acorde los valores de *arousal* y *valence*.

Debido la crecimiento exponencial de nueva musica, es necesario crear nuevas maneras de organizar la musica por emoción. Un manera de afrontar este problema es llamado clasificación de emoción de musica (MEC) que divide las emociones en clases y aplica métodos de machine learning en características de audio para reconocer las emociones que son transmitidas en la frecuencia de la musica. Debido a la separación semántica, el progreso de modelos mono-modales ha sido estancado. Para complementar el resultado de estos modelos se ha estudiado el impacto de los modelos multi-modales, donde se emplea la letra de las canciones ya que es semánticamente rica y expresiva y tiene un profundo impacto en la percepción humana en la musica. Es fácil para nosotros decir si una letra expresa amor, tristeza, felicidad o algo mas, por lo que incorporar la letra en un modelo multimodal ayudara al análisis de la emoción de la musica.

II. Antecedentes

II.1. Recuperación de Información Musical (MIR)

La recuperación de información musical es un campo de investigación dedicado a desarrollar métodos y técnicas para buscar, organizar y acceder a información relacionada con la musica. Específicamente el libro titulado "Fundamentals of Music Processing" de Müller, M., 2015 [7] proporciona una visión general exhaustiva de los conpetos fundamentales y técnicas utilizadas en el procesamiento de música, donde se incluyen temas en MIR.

II.2. Análisis de Emoción y Sentimiento en Música

El análisis de emoción y sentimiento en música (MER) ha sido un campo de estudio ampliamente explorado en la investigación de procesamiento de señales de audio y minería de datos musicales. Los estudios de MER se centran en comprender y extraer información emocional y afectiva de la música utilizando diversas técnicas computacionales y modelos de aprendizaje automático. En el artículo "Automatic Mood Classification of Music Content" (P. Bogdanov, et al., 2013) [1] se ofrece una recompilación de la visión general exhaustiva de las metodologías y técnicas utilizadas en la clasificación automática de la emoción y sentimiento musical.

II.3. Perspectiva Multimodal y Extracción de Características Multimodales

El trabajo en información automática o métodos predictivos de música empieza con el desarrollo de la música digital. el estudio "The annual Music Information Research Evaluation eXchange (MIREX)" incluyó por primera vez el estado de ánimo musical como una tarea de clasificación, encontrando que la etiquetación de la verdad fundamental y el juicio humano del estado de ánimo juegan un papel crítico en esta tarea [3]. Se utilizaron una variedad de características, desde características espectrales de bajo nivel (señal de frecuencia cruda, como centroides espectrales y desviación estándar) [9] hasta características de nivel superior como "danceability", ritmo o timbre, que se pueden extraer de características de bajo nivel y se describen más cercanas a la percepción humana.

En el estudio de Laurier et al. (2008) [6] se implementó las características de las letras de canciones para incluir recuentos de n-grams, las cuales son características estilísticas y la anotación de sentimientos ANEW [2] y encontraron que las letras tienden a tener mejor desempeño que las características auditivas en tareas de clasificación y regresión musical, aunque los mejores resultados se lograron al combinar las características. Por otro lado, han habido muchas propuestas de modelos Transformes que tienen como objetivo manejar características de texto, numéricas y categóricas. En el caso de los Transformers pre-entrenados como ViLBERT y VLBERT [8] siguen principalmente el mismo enfoque que el modelo original BERT, pero tratan tokens adicionales en la entrada, lo que ofrece una visión interesante sobre cómo combinar características categóricas y numéricas.

La relación de las características de las dos variables predichas se describe el arousal relacionado con el tempo (rápido/lento), pitch (alto/bajo), loudness y timbre. Así como la valencia se relaciona con el mode (major/minor) y la armonía [4]. Estás variables consisten en puntuaciones positivas, negativas, neutrales y compuestas, como se muestra en la figura 1:

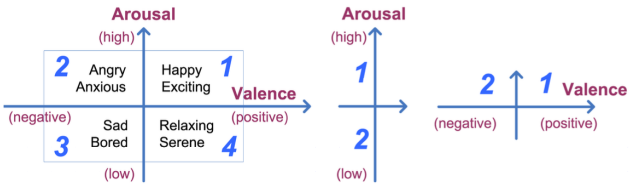

Figura 1: Scores para las variables Arousal y Valencia.

III. Metodología

En este trabajo se utilizan modelos Unimodales con características de música o letra de manera excluyente y con el fin de visualizar la mejora de la combinación de caraterísticas se estara utilizando tanbién un modelo de Multimodal-Toolkit que utilizan Transformes para situaciones que incluyen texto y datos tabulares. Estos diseños permiten incluir fácilmente modelos Transformer que utilizan un pipeline de entrenamiento con varias características e integran los modelos entrenados de Hugging Face, en este caso se usa el Transformer DistilBert, que proporciona un módulo de preprocesamiento de datos para características de texto, categóricas y numéricas. Se muestra una visión general del modelo en la figura 2:

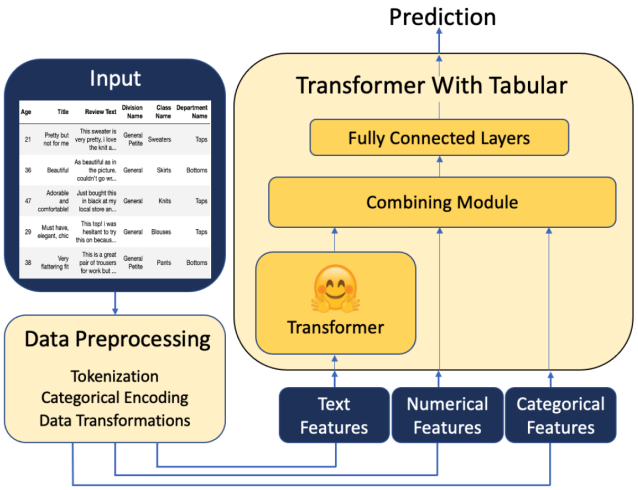

Figura 2: Estructura de Multimodal-Toolkit. La información procesada como salida tuene características de texto, numéricas y categóricas que son implementadas como entradas al Transformador de Hugging Face [5].

III.1. Base de datos

La base de datos consiste en la extracción de características de canciones de la base de datos de Deezer Mood Detection Dataset (DMDD) que tiene valores para 13,445 canciones con 11 características de audio que son obtenidas mediante el API de Spotify. Las características consisten en:

- **Acousticness:** medida numérica de 0.0 a 1.0 si es pista es acústica, 1.0 representa una alta confianza de que la pista es acústica.

- **Danceability:** describe que tan adecuado es una pista para bailar basado en una combinación de elementos musicales, donde se incluyen el tempo, ritmo, intensidad del compás, y uniformidad global. Un valor de 0.0 es menos bailable y de 1.0 es muy bailable.
- **Energy:** representa una medida de 0.0 a 1.0 de intensidad y actividad. Las pista energéticas se sienten rápidas y ruidosas. Los atributos que incluyen son rango dinámico, ruido percibido, timbre, velocidad de inicio, y entropía general.
- **Instrumentalness:** Predice si el audio contiene o no voz. Un valor mas cercano a un valor de 1.0 es instrumental.
- **Key:** valor categorico que indica la representación de notas musicales en notación de clases de tono. Por ejemplo, 0 representa la nota C, 1 representa C#/Db, 2 representa D y así sucesivamente. Si no se detecta ninguna tonalidad, el valor sería -1.
- **Liveness:** detecta la presencia de una audiencia en la grabación. Valores altos representan una probabilidad alta de que la pista fue ejecutado en vivo.
- **Loudness:** se refiere al nivel promedio de volumen a lo largo de toda la pista. Es la cualidad de un sonido que se correlaciona principalmente con la amplitud. Sus valores oscilan entre -60 y 0 dB. Esta medida proporciona una manera estandarizada de evaluar y comparar el volumen entre distintas grabaciones.
- **Mode:** indica la modalidad (major or minor) de una pista, el tipo de escala del cual su contenido melódico es derivado. 1 representa major y 0 minor.
- **Speechiness:** detecta la presencia de palabras habladas en la pista. Cuanto más parecido a la voz humana sea la grabación, más cercano será el valor a 1.0. Valores entre 0.33 y 0.66 describen pistas que pueden contener tanto música como palabras habladas.
- **Tempo:** es la velocidad o ritmo de una pieza musical y se mide en beats por minuto (BPM). Es la velocidad a la que se suceden los beats en una canción.
- **Valence:** medida de 0.0 a 1.0 que describe la positividad musical transmitida por una pista. Las pistas con un valor alto de valencia suenan más positivas, alegres, animadas o eufóricas, mientras que aquellas con un valor bajo de valencia suenan más negativas, tristes, deprimidas o enojadas. Es una forma de medir el tono emocional general que transmite una canción, yendo desde emociones positivas hasta emociones negativas, representadas en el rango de valores de valencia
- **Lyrics:** atributo textual que contiene la letra de la canción. En la sección **III.2** se describe el proceso de extracción de características para este elemento.
- **y_arousal:** refleja el nivel de excitación o activación emocional inducido por la canción, abarcando desde estados más calmados hasta emociones intensas y enérgicas.
- **y_valence:** captura la valencia emocional, que abarca el espectro desde emociones negativas (tristeza, melancolía) hasta emociones positivas (felicidad, alegría).

III.2. Extracción de Características de Letras de Canciones

Para capturar las características semánticas de las letras de las canciones, se implementó un método de extracción de características basado en el modelo de lenguaje preentrenado DistilBERT. Este método se diseñó con el objetivo de obtener representaciones densas de las letras que pudieran utilizarse en tareas posteriores, como clasificación de género musical o análisis de sentimiento.

III.2.1. Configuración del Modelo y Tokenizador

Se empleó la arquitectura "distilbert-base-uncased"para el modelo DistilBERT, conocida por su eficiencia computacional sin sacrificar el rendimiento. El tokenizador asociado a esta arquitectura fue utilizado para dividir las letras de las canciones en unidades discretas, preservando la estructura y el contexto de las palabras.

```
    model_name = 'distilbert-base-uncased'
    tokenizer = DistilBertTokenizer.from_pretrained(model_name)
    model = DistilBertModel.from_pretrained(model_name)
```

III.2.2. Tokenización y Obtención de Representaciones Ocultas

Las letras de las canciones fueron tokenizadas utilizando el tokenizador previamente cargado. Posteriormente, se obtuvieron las representaciones ocultas de la capa intermedia del modelo DistilBERT mediante el paso de las secuencias tokenizadas al modelo.

```
    inputs = tokenizer(song_lyrics, return_tensors='pt', truncation=True, padding=
    True)
    outputs = model(**inputs)
    hidden_states = outputs.last_hidden_state
```

Los atributos de *truncation* y *padding* nos ayudan a estandarizar las dimensiones del texto para poder tokenizarlas.

III.2.3. Obtención de una Representación General

Para obtener una representación general de la letra de la canción, se calculó el promedio de las representaciones de todas las palabras en la capa intermedia. Este enfoque permitió condensar la información semántica de la letra en una representación vectorial de dimensiones reducidas.

```
    avg_representation = torch.mean(hidden_states, dim=1).detach().numpy()
```

Este proceso resulta en un vector de características que resume las propiedades semánticas de la letra de la canción y puede ser utilizado como entrada para tareas subsiguientes de análisis y clasificación.

III.3. Arquitectura de Red Neuronal

Para abordar el problema de modelado en nuestro estudio, se diseñó e implementó una arquitectura de red neuronal personalizada denominada CustomNN. Esta red se concibió con el propósito de capturar patrones complejos en los datos de entrada y generar salidas relevantes para la tarea específica.

III.3.1. Estructura de la Red Neuronal

La arquitectura de CustomNN consta de capas de entrada, capas ocultas y una capa de salida lineal. La configuración de la red se define mediante los siguientes elementos:

Capa de Entrada

```
    self.input_layer = nn.Linear(input_size, hidden_sizes[0])
```

La capa de entrada transforma las características iniciales de las muestras de entrada al espacio de representación de la primera capa oculta.

Capas Ocultas

```
    self.hidden_layers = nn.ModuleList([
        nn.Sequential(
            nn.Linear(hidden_sizes[i], hidden_sizes[i+1]),
            nn.Sigmoid()
        )
        for i in range(len(hidden_sizes) - 1)
    ])
```

Las capas ocultas son modeladas como una secuencia de capas lineales seguidas de funciones de activación sigmoide. Cada capa oculta captura y procesa información relevante a medida que se profundiza en la red.

Capa de Salida Lineal

```
    self.output_layer = nn.Linear(hidden_sizes[-1], output_size)
```

La capa de salida lineal genera las predicciones finales basadas en las representaciones aprendidas por las capas anteriores.

III.3.2. Función de Propagación Hacia Adelante

La función *'forward'* define cómo los datos se propagan a través de la red:

```
    def forward(self, x):
    x = self.input_layer(x)

    for hidden_layer in self.hidden_layers:
        x = hidden_layer(x)

    x = self.output_layer(x)
    return x
```

Esta función implementa la propagación hacia adelante, donde las entradas son transformadas capa por capa hasta que se obtiene la salida final. La función de activación sigmoide en las capas ocultas introduce no linealidades en el modelo, permitiendo la captura de patrones más complejos.

IV. Resultados

Primeramente se muestran los resultados para el modelo unimodal con características únicamente de la letra de la canción.

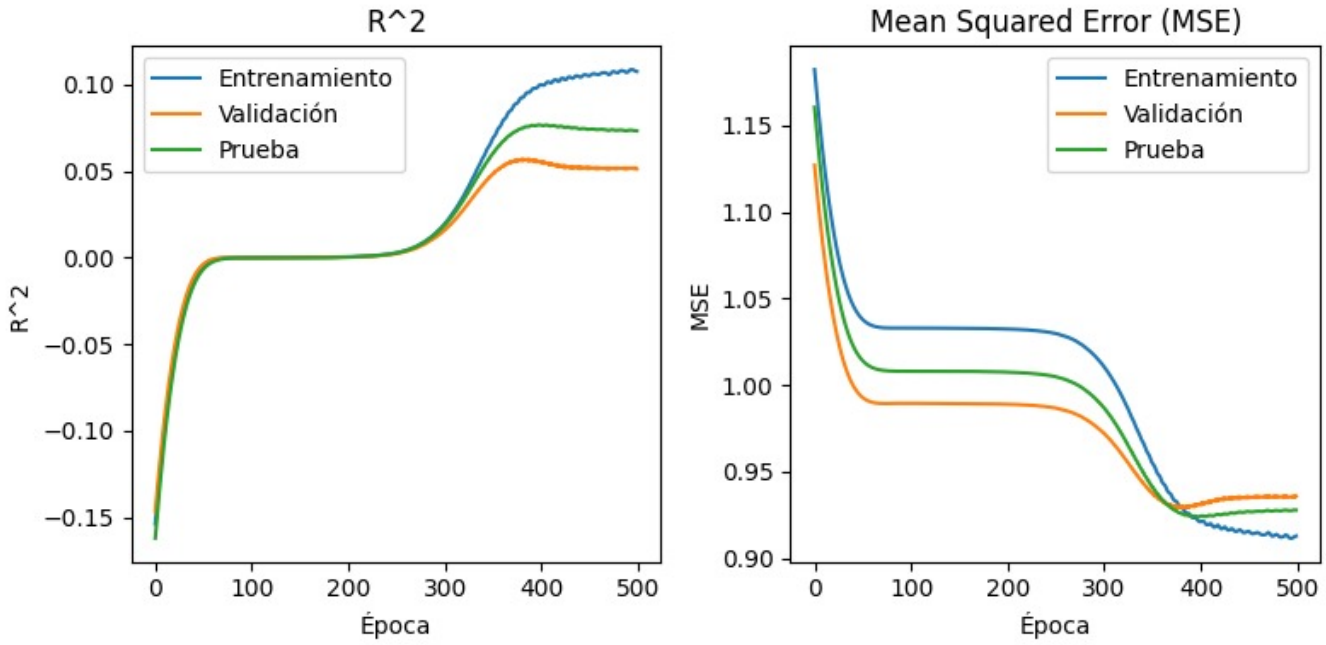

Figura 3: Perdida y Ajuste de Modelo con características de letra.

Los resultados de la fase de prueba para el modelo unimodal, utilizando únicamente características derivadas de las letras de las canciones, revelan un coeficiente de determinación R^2 de 0.07. Este valor sugiere que el modelo explica solo un 7 % de la variabilidad en los datos de prueba, indicando que las características extraídas de las letras por sí solas pueden no ser suficientes para capturar la complejidad subyacente en la tarea de predicción. Es crucial explorar estrategias adicionales, como la inclusión de información contextual o la expansión de las características utilizadas, para mejorar la capacidad predictiva de este modelo unimodal.

Además, el Error Cuadrático Medio (MSE) de 0.92 indica que las predicciones del modelo se desvían considerablemente de los valores reales en el conjunto de prueba. Este resultado sugiere que, aunque las características de las letras se han utilizado como único input, su capacidad para explicar la variabilidad en la variable objetivo puede ser limitada. Es recomendable considerar enfoques más sofisticados, como la combinación de múltiples modalidades o la incorporación de información contextual, para mejorar la capacidad predictiva del modelo en el contexto específico de las letras de las canciones. En resumen, los resultados actuales resaltan la necesidad de estrategias adicionales para fortalecer la capacidad de predicción del modelo unimodal basado en características de letras.

Ahora bien, para el modelo unimodal con características de la letra se obtuvieron los siguientes resultados.

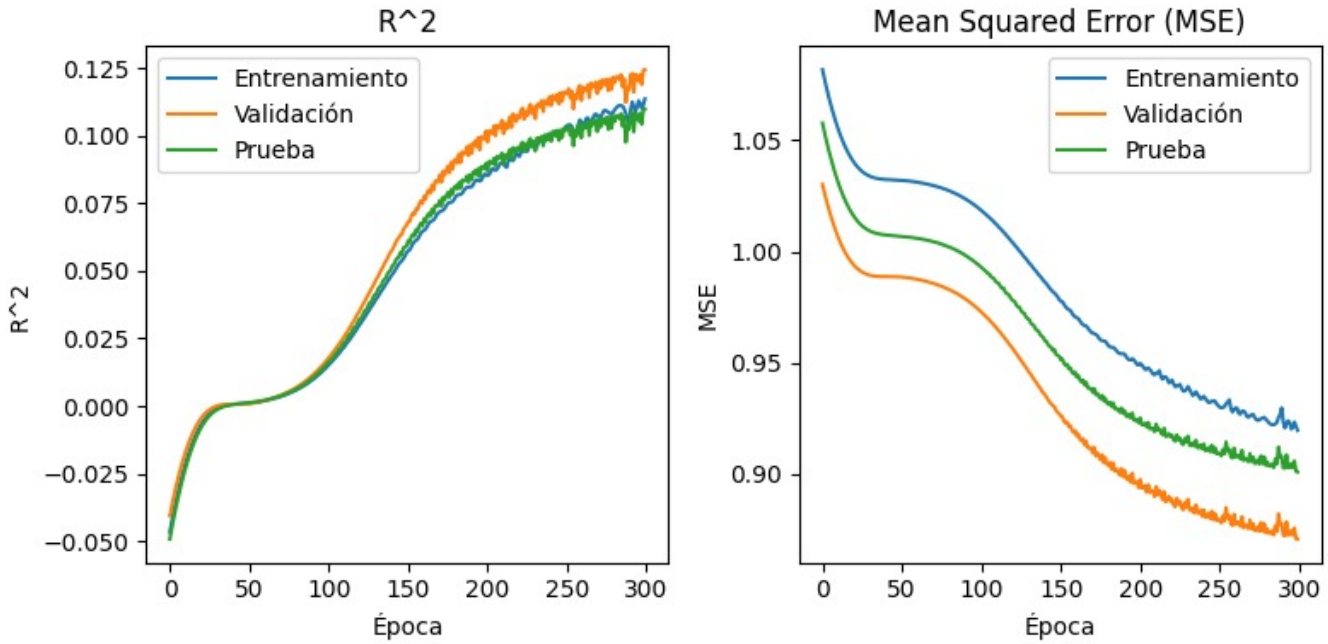

Figura 4: Perdida y Ajuste de Modelo con características de música.

Los resultados obtenidos en la fase de prueba para el modelo unimodal, centrado exclusivamente en características derivadas de la música de las canciones, muestran un coeficiente de determinación (R^2) de 0.1. Aunque este valor indica una capacidad limitada del modelo para explicar la variabilidad en los datos de prueba, sugiere una cierta capacidad predictiva. Sin embargo, la proporción relativamente baja de variabilidad explicada (10 %) destaca la necesidad de considerar estrategias complementarias o la inclusión de características adicionales para mejorar la precisión del modelo en la predicción de la variable objetivo.

Además, el Error Cuadrático Medio (MSE) de 0.9 indica que las predicciones del modelo están, en promedio, cercanas a los valores reales en el conjunto de prueba. Aunque la capacidad predictiva puede considerarse moderada, es importante explorar la inclusión de información contextual o la expansión de las características musicales para abordar posibles limitaciones en la capacidad explicativa del modelo. En resumen, los resultados sugieren un rendimiento prometedor del modelo unimodal basado en características de la música, pero se destaca la oportunidad de mejoras adicionales para potenciar su capacidad predictiva y explicativa.

Finalmente, el modelo Multimodal que hizo uso de ambos tipos de características nos arrojo los resultados presentados a continuación:

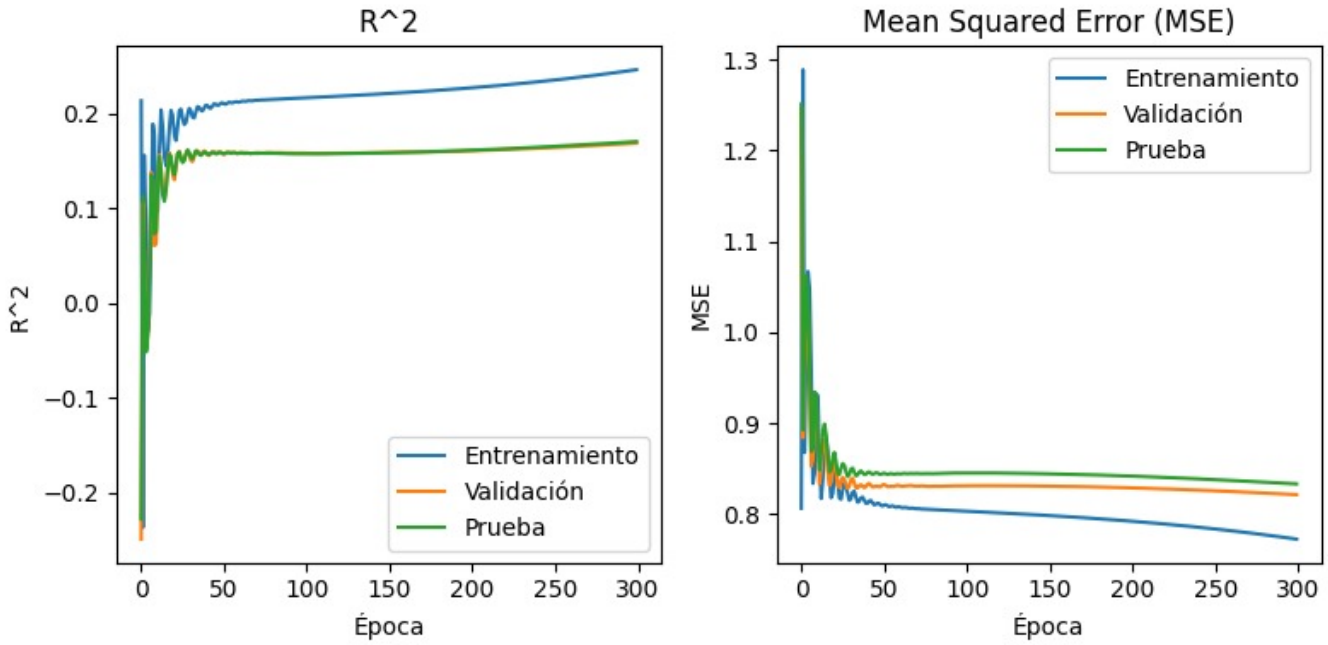

Figura 5: Perdida y Ajuste de Modelo Multimodal.

Los resultados obtenidos en la fase de prueba para el modelo multimodal revelan un coeficiente de determinación R^2 de 0.17, indicando que el modelo explica aproximadamente el 17 % de la variabilidad en los datos de prueba. Este R^2 relativamente bajo sugiere que el modelo podría no estar capturando completamente las complejidades inherentes a la relación entre las modalidades. Es esencial considerar posibles mejoras en la arquitectura del modelo o la inclusión de características adicionales para fortalecer su capacidad predictiva.

Además, el Error Cuadrático Medio (MSE) de 0.83 destaca que, en promedio, las predicciones del modelo difieren significativamente de los valores reales en el conjunto de prueba. Este valor relativamente alto indica que el modelo podría beneficiarse de ajustes adicionales para mejorar la precisión y la calidad de las predicciones. En resumen, aunque el modelo multimodal presenta un rendimiento predictivo, las métricas sugieren oportunidades para refinamiento y ajuste, lo que puede conducir a una mejora sustancial en su capacidad para abordar la complejidad de la tarea multimodal.

V. Conclusiones y Discusión

En este estudio, se abordó el análisis de emoción y sentimiento en la música desde una perspectiva multimodal, utilizando características de letra y música. A continuación, se presentan algunas conclusiones y reflexiones sobre los resultados obtenidos:

V.1. Rendimiento de Modelos Unimodales

Se evaluaron modelos unimodales utilizando únicamente características de letra o música. Los resultados indican que, aunque estos modelos muestran cierta capacidad predictiva, la variabilidad explicada es limitada. Se observó un coeficiente de determinación (R^2) bajo y un Error Cuadrático Medio (MSE) que

sugiere desviaciones significativas entre las predicciones y los valores reales. Esto resalta la complejidad de la tarea y la necesidad de considerar enfoques más sofisticados.

V.2. Modelo Multimodal

El modelo multimodal, que incorpora tanto características de letra como de música, demostró un rendimiento prometedor. Sin embargo, el R^2 y el MSE indican que aún hay margen para mejoras. La capacidad de explicar la variabilidad en los datos de prueba podría beneficiarse de ajustes adicionales en la arquitectura del modelo o la inclusión de más características.

V.3. Oportunidades de Mejora

Es crucial considerar estrategias para mejorar la capacidad predictiva del modelo multimodal. Esto podría incluir la exploración de características adicionales, la optimización de hiperparámetros o la consideración de arquitecturas más complejas. Además, la incorporación de técnicas de preprocesamiento avanzadas o el uso de modelos preentrenados podrían contribuir a la mejora del rendimiento.

V.4. Limitaciones y Consideraciones Futuras

Es importante reconocer las limitaciones del estudio, como la naturaleza subjetiva de las etiquetas de emoción, la cantidad de datos y la variabilidad intrínseca en las preferencias musicales. Las futuras investigaciones podrían abordar estos aspectos y considerar la expansión del conjunto de datos para obtener una comprensión más completa de las complejidades emocionales en la música.

En conclusión, este estudio proporciona una visión inicial del análisis multimodal de emoción y sentimiento en la música. Aunque se lograron avances, hay oportunidades significativas para el refinamiento de los modelos y la exploración de enfoques más avanzados en futuras investigaciones.

Referencias

[1] D. Bogdanov et al. From music similarity to music recommendation: Computational approaches based on audio features and metadata. 2013.

[2] M. M. Bradley and P. J. Lang. Affective norms for english words (anew): Instruction manual and affective ratings. Technical report, Technical report C-1, the center for research in psychophysiology . . . , 1999.

[3] X. Downie, C. Laurier, and M. Ehmann. The 2007 mirex audio mood classification task: Lessons learned. In *Proc. 9th Int. Conf. Music Inf. Retrieval*, pages 462–467, 2008.

[4] A. Gabrielsson and E. Lindström. The influence of musical structure on emotional expression. 2001.

[5] K. Gu and A. Budhkar. A package for learning on tabular and text data with transformers. In *Proceedings of the Third Workshop on Multimodal Artificial Intelligence*, pages 69–73, 2021.

[6] C. Laurier, J. Grivolla, and P. Herrera. Multimodal music mood classification using audio and lyrics. In *2008 seventh international conference on machine learning and applications*, pages 688–693. IEEE, 2008.

[7] M. Müller. *Fundamentals of Music Processing*. Springer, 2015.

[8] W. Su, X. Zhu, Y. Cao, B. Li, L. Lu, F. Wei, and J. Dai. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*, 2019.

[9] Y.-H. Yang, Y.-C. Lin, H.-T. Cheng, I.-B. Liao, Y.-C. Ho, and H. H. Chen. Toward multi-modal music emotion classification. In *Advances in Multimedia Information Processing-PCM 2008: 9th Pacific Rim Conference on Multimedia, Tainan, Taiwan, December 9-13, 2008. Proceedings 9*, pages 70–79. Springer, 2008.