ISSN 1870-4069

Identificación de patrones lingüísticos, temáticos, emocionales e intención en los corridos y la música regional mexicana

José Eder Martínez-Martínez [1], Yessenia Díaz-Álvarez [1], Tláloc Humberto Vergara-Morales [1], Víctor Manuel Millán-Aguilar [2], Gabriel González-Serna [1], Noé Alejandro Castro-Sánchez [1]

[1] Tecnológico Nacional de México, campus CENIDET, México

[2] Tecnológico Nacional de México, campus Zacatepec, México

{m24ce005, d24ce109, m23ce088, gabriel.gs, noe.cs}
@cenidet.tecnm.mx, L20091134@zacatepec.tecmm.mx

Resumen. La música regional mexicana, especialmente los corridos, ha experimentado una evolución significativa en su contenido lírico, reflejando influencias culturales, sociales y digitales. Este estudio realiza un análisis computacional integral de los corridos, combinando modelado de tópicos, análisis de sentimientos y clasificación de intención. A través del Reconocimiento de Entidades Nombradas (NER), se identifican términos recurrentes y referencias geográficas que evidencian la influencia de la migración y la globalización. Además, mediante BERTopic, se extraen los temas predominantes y se analizan las tendencias temáticas. Para evaluar la carga emocional de las letras, se emplea una variante del modelo RoBERTa-base-bne, permitiendo la detección automática de sentimientos en los textos, analizando los cambios lingüísticos y patrones recurrentes. Finalmente, empleando BART-large-MNLI, se clasifican las canciones según si glorifican, desprestigian, o simplemente describen actividades delictivas. Los resultados proporcionan una visión profunda de los mensajes transmitidos a través de los corridos y sus implicaciones culturales y sociales.

Palabras clave: Procesamiento de lenguaje natural, modelado de tópicos, corridos mexicanos, análisis de sentimientos y emociones, clasificación de intención.

Identifying Linguistic, Thematic, Emotional and Intent Patterns in Corridos and Mexican Regional Music

Abstract. Mexican Regional Music, particularly “corridos”, have experimented a major evolution within its lyrical content, reflecting new cultural, social and digital influences. This study conducts a comprehensive computational analysis of “corridos”, combining topic modeling, sentiment analysis, and intent classification. Through Named Identity Recognition (NER), recurring terms and

geographical references are identified, evidencing the influence of migration and globalization. Additionally, through BERTopic, the predominant themes are extracted and the thematic trends are analyzed. To evaluate the emotional load of the lyrics, a variant of the RoBERTa-base-bne model is used, allowing the automatic detection os sentiments in the texts, analyzing linguistic changes and recurring patterns. Finally, using BART-large-MNLI, the songs are classified according to whether the y glorify, vilify or simply describe criminal activities. The results provide an in-depth view of the messages conveyed through "corridos" and their cultural and social implications.

Keywords: Natural language processing, topic modeling, mexican corridos, sentiment and emotion analysis, intent classification.

1. Introducción

Los corridos mexicanos han sido un medio de narración que ha evolucionado junto con los cambios sociales y culturales, abordando desde historias de héroes revolucionarios hasta figuras del crimen organizado. Su contenido refleja influencias de la globalización y la era digital, transformando su narrativa para conectar con nuevas audiencias.

Este estudio emplea técnicas de Procesamiento de Lenguaje Natural (PLN) para analizar la estructura y evolución temática de los corridos y la música regional mexicana. Mediante BERTopic, se identifican los temas predominantes, mientras que el Reconocimiento de Entidades Nombradas (NER) permite detectar patrones lingüísticos clave, como nombres propios, lugares y términos recurrentes que reflejan las principales narrativas del género. Además, el análisis de sentimientos con RoBERTa-base-bne permite evaluar la carga emocional de las letras, proporcionando una perspectiva más profunda sobre su impacto.

Por último, la clasificación de intención con BART-large-MNLI ayuda a distinguir entre canciones que glorifican el crimen y aquellas que simplemente lo narran. A través de esta combinación de técnicas, se busca comprender mejor el papel de los corridos en la música y la sociedad contemporánea.

2. Trabajos relacionados

El Reconocimiento de Entidades Nombradas (NER) y el análisis de frecuencia de términos han sido utilizados para estudiar patrones temáticos en textos musicales. Investigaciones previas han aplicado estas técnicas en distintos enfoques [1, 2], pero su uso en los corridos y la música regional mexicana sigue siendo limitado. El análisis de frecuencia léxica ha demostrado ser útil para identificar tendencias en la música [3], permitiendo comprender cambios en el contenido lírico. Este estudio combina NER y análisis de términos para explorar la evolución temática de los corridos, aportando una visión cuantitativa sobre sus principales narrativas.

Existen diversos estudios sobre el impacto de los géneros musicales en la sociedad y la detección de temáticas en letras de canciones. Investigaciones previas han aplicado técnicas de modelado de tópicos para analizar narrativas en la música popular y su

relación con estereotipos de género y violencia. Sin embargo, pocos estudios se han centrado específicamente en los corridos. El modelado de temas agrupa documentos para resumirlos o clasificarlos. Al aplicarse a las letras de las canciones, proporciona un método eficaz para identificar temas recurrentes [4, 5]. Algoritmos de modelado de temas como BERTopic se han utilizado previamente en la investigación de género y ciencias sociales, y Wickham [6] utilizó este algoritmo para estudiar las expectativas de género en las redes sociales y su impacto en la ideación suicida.

Algunos estudios han explorado cómo estas canciones influyen en la percepción de la violencia y el crimen en la sociedad, mientras que otros han examinado su papel como una forma de protesta o resistencia cultural [7, 8]. Además, en el campo de la lingüística y el análisis de contenido, el Procesamiento de Lenguaje Natural (PLN) ha permitido analizar los temas y el sentimiento de las letras de manera automática, proporcionando una perspectiva sobre las emociones predominantes y la narrativa detrás de estos temas [9].

En cuanto a la clasificación temática e inferencia de intención en los corridos, BART-large-MNLI ha sido utilizado para tareas de categorización de texto sin necesidad de entrenamiento adicional. Biswas [10] desarrolló una API utilizando Flask que integra este modelo, facilitando su aplicación en entornos de producción. Un modelo derivado, BART-Large-MNLI-Yahoo-Answers, ha sido ajustado específicamente para la categorización de temas en el conjunto de datos de Yahoo Answers, demostrando una alta precisión en la clasificación de textos, incluso en categorías no vistas durante el entrenamiento [11]. Además, se han discutido implementaciones prácticas de BART-large-MNLI en foros como Stack Overflow, abordando su despliegue en producción y su integración en interfaces de usuario [12]. Estas discusiones reflejan el creciente interés por el uso de modelos de inferencia de lenguaje natural en aplicaciones prácticas.

3. Metodología

En esta sección se detallan los procedimientos utilizados para recolectar, procesar y analizar los datos. Que se pueden encontrar en el siguiente enlace https://github.com/JoseEderMartinezMartinez/AnalisisComputacionalCorridosyMusicaRegionalMexicana

a. Recopilación de datos

Para la compilación del corpus, se implementó un sistema de web scraping con Scrapy, obteniendo letras de canciones de los géneros corrido y música regional mexicana desde la plataforma letras.com.

El proceso de scraping se estructuró en varias etapas:

1. Obtención de artistas por género: Se utilizó un archivo maestro en formato JSON con una lista de artistas de los géneros objetivo.

2. Extracción de enlaces de canciones: Se recorrieron las páginas de cada artista para recolectar los enlaces de sus canciones.

3. Descarga de letras: Se accedió a cada enlace de canción y se extrajo el contenido textual de las letras.

Para evitar restricciones en la extracción masiva, se integró un sistema de proxies dinámicos, obtenidos y validados previamente desde free-proxy-list.net. Adicionalmente, se implementaron filtros para eliminar letras duplicadas y registros incompletos.

Este método permitió la construcción de un corpus diverso y representativo de los géneros analizados, la identificación de patrones en las letras de corridos y música regional mexicana.

b. Limpieza de datos

Una vez recopiladas las letras de canciones, se llevó a cabo un proceso de limpieza para mejorar la calidad del corpus y asegurar que los datos fueran adecuados para el análisis de tópicos. Este preprocesamiento incluyó varias etapas, con el objetivo de eliminar ruido y normalizar los textos.

Eliminación de Datos No Relevantes. Se aplicaron filtros para descartar aquellas canciones que presentaban características no deseadas:

- Letras vacías o compuestas únicamente por espacios en blanco.
- Textos que contenían exclusivamente caracteres especiales o números sin contenido léxico significativo.
- Canciones con errores en la estructura del archivo JSON o con registros incompletos.
- Canciones escritas en idiomas distintos al español.

Normalización y Corrección Ortográfica. Para mejorar la coherencia del corpus y reducir la variabilidad léxica, se implementaron los siguientes pasos:

- Conversión a minúsculas para uniformar los datos y evitar diferencias artificiales entre palabras.
- Eliminación de signos de puntuación y caracteres especiales, a excepción de aquellos relevantes para la estructura gramatical.
- Corrección ortográfica automática utilizando la biblioteca pyspellchecker [15], con un diccionario de palabras en español, asegurando que las palabras estuvieran correctamente escritas.

Estandarización de Expresiones. Dado que las letras de los corridos suelen incluir palabras coloquiales o variantes ortográficas propias del género, se compararon las palabras del corpus con un diccionario de términos en español para detectar y corregir términos no estándar o con errores tipográficos frecuentes.

El corpus final se dividió en dos conjuntos de datos:

Tabla 1. Características de los Datasets.

Género	Número total de canciones	Número de artistas únicos
Corridos	2,496	300
Regional	2,944	286

Pseudocódigo 1. Extracción de entidades.

```
FUNCIÓN obtener_terminos_ner(texto)
  doc ← aplicar_modelo_NER(texto)  // Procesa el texto con el modelo NER
  terminos ← lista_vacía  // Lista para almacenar los términos extraídos
  PARA cada entidad ent EN doc.entidades
    SI contar_palabras(ent.texto) > 1 ENTONCES
      agregar_a_lista(terminos, convertir_a_minusculas(ent.texto))
    FIN SI
  FIN PARA
  RETORNAR terminos
FIN FUNCIÓN
```

Estas transformaciones permitieron construir un conjunto de datos limpio y estructurado, optimizando la detección de patrones temáticos en los corridos y la música regional.

c. Extracción de términos con NER

Para obtener los términos clave, se implementó la función "obtener_terminos_ner", con el objetivo de procesar el texto utilizando el modelo "es_core_news_md" de la biblioteca de "spaCy" y extraer las entidades compuestas por dos o más palabras. Se presenta su estructura en el **Pseudocódigo 1** para facilitar la comprensión del proceso.

Esta función primero procesa el texto con el modelo de spaCy para realizar el Reconocimiento de Entidades Nombradas (NER), luego filtra las entidades detectadas asegurando que solo se incluyan aquellas con más de una palabra, y finalmente las convierte a minúsculas antes de devolver la lista de términos extraídos. Este enfoque permite identificar palabras clave que reflejan la estructura y el contenido del dataset analizado.

d. Análisis de frecuencia de términos

Tras la extracción de términos **NER**, se aplicó un análisis de frecuencia para identificar las palabras más comunes dentro del dataset. Para ello, se utilizó el **Pseudocódigo 2**.

Pseudocódigo 2. Análisis de frecuencia.

```
INICIO
// Descomponer listas en filas individuales
terminos ← descomponer_listas(df['terminos'])
// Contar la frecuencia de cada término
conteo ← contar_frecuencia(terminos)
// Convertir conteo a una tabla para mejor visualización
df_frecuencia ← crear_tabla(conteo, columnas=['Término', 'Frecuencia'])
// Ordenar la tabla por frecuencia en orden descendente
df_frecuencia ← ordenar_por_columna(df_frecuencia, 'Frecuencia', descendente=True)
// Mostrar los 50 términos más frecuentes
imprimir(primeros_n_elementos(df_frecuencia, 50))
FIN
```

Este procedimiento permite descomponer las listas de términos en filas individuales, contar su frecuencia y estructurar los datos en una tabla ordenada. Posteriormente, se filtran los términos con una frecuencia mínima de 1 y se seleccionan los 50 más comunes para su análisis.

e. Preprocesamiento de datos para detección de tópicos

BERTopic, desarrollado por Grootendorst [13], es un modelo de detección de tópicos basado en técnicas de aprendizaje profundo que combina embeddings generados por modelos de lenguaje preentrenados con métodos de reducción de dimensionalidad y clustering.

Su procedimiento consta de cuatro etapas principales:

Generación de incrustaciones de texto. Se utilizan modelos de lenguaje preentrenados para transformar cada documento (letra de canción) en una representación vectorial en el espacio semántico. Para este estudio, se empleó el modelo MiniLM-L6-V2, optimizado para tareas de representación de textos.

Reducción de dimensionalidad. Dado que los embeddings suelen tener una dimensionalidad alta, se aplica UMAP (Uniform Manifold Approximation and Projection) para reducir la complejidad del espacio y preservar relaciones semánticas clave entre documentos.

Agrupamiento de documentos. Los embeddings reducidos son agrupados mediante HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise), lo que permite identificar grupos de canciones con características temáticas similares. Este enfoque permite gestionar la variabilidad de los textos sin necesidad de especificar un número fijo de tópicos.

Extracción de términos representativos. Para obtener una representación semántica interpretable de cada tema, se utiliza c-TF-IDF (class-based Term Frequency-Inverse Document Frequency), que permite identificar las palabras más relevantes dentro de cada clúster, diferenciándolas de otros tópicos en el datasets.

f. Análisis de sentimientos

El modelo utilizado para realizar el análisis de sentimientos fue edumunozsala/roberta_bne_sentiment_analysis_es. Este modelo se basa en la arquitectura RoBERTa-base-bne, una variante de RoBERTa preentrenada con el mayor corpus en español conocido hasta la fecha, con un total de 570 GB. Este corpus fue desarrollado por la Biblioteca Nacional de España [14].

El modelo fue afinado específicamente para análisis de sentimientos en español utilizando un conjunto de datos de aproximadamente 50,000 reseñas de películas en español. Este conjunto de datos está equilibrado y proporciona cada reseña en inglés y español, junto con las etiquetas en ambos idiomas.

g. Análisis de emociones

Para el análisis de emociones se utilizó el modelo bhadresh-savani/distilbert-base-uncased-emotion, el cual es una versión de DistilBERT, un modelo de lenguaje que utiliza destilación de conocimiento durante la fase de preentrenamiento para reducir el tamaño de BERT en un 40%, manteniendo el 97% de su capacidad de comprensión del lenguaje. Este modelo es más pequeño y rápido que BERT y otros modelos basados en BERT.

La arquitectura de DistilBERT se basa en la de BERT, pero con algunas modificaciones para hacerlo más eficiente.

El modelo distilbert-base-uncased-emotion ha sido ajustado específicamente para la clasificación de emociones utilizando el conjunto de datos de emociones de Twitter.

h. Análisis de intención

Para este análisis se empleó BART, el cual es un modelo de transformador secuencia a secuencia que combina las características de un codificador bidireccional, similar a BERT, y un decodificador autorregresivo, similar a GPT. Esta arquitectura le permite manejar tareas de generación y comprensión de lenguaje natural de manera eficiente. El proceso de preentrenamiento de BART implica dos etapas principales:

1. Corrupción del texto: Se aplica una función de ruido al texto original para corromperlo, lo que puede incluir la eliminación de tokens, permutaciones de oraciones, entre otros métodos.

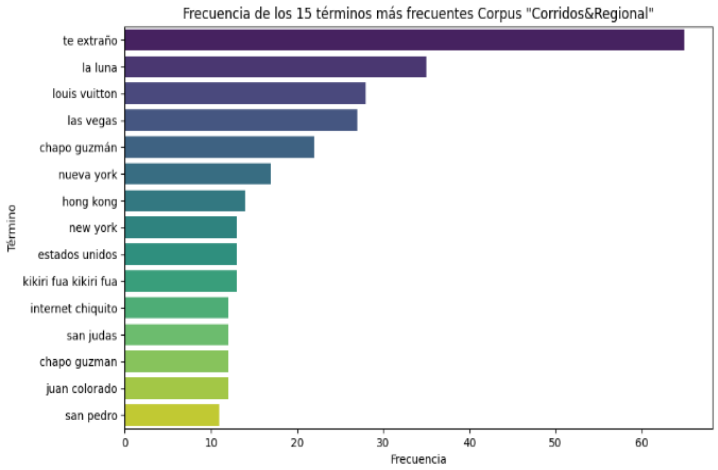

Fig. 1. Los 15 términos más frecuentes en el dataset "Corridos&Regional".

2. Reconstrucción del texto: El modelo aprende a reconstruir el texto original a partir de la versión corrompida, desarrollando así una comprensión profunda de la estructura y el significado del lenguaje.

Para este trabajo se especificaron las etiquetas a detectar por el modelo: “Glorificar el narcotráfico”, “Describir el narcotráfico” y “Desprestigiar el narcotráfico”.

4. Resultados y experimentos

En esta sección se muestran los resultados de los experimentos realizados en cada modelo previamente mencionado.

a. Extracción de términos con NER

Para este experimento se realizó el análisis de los dos datasets de música regional y corridos y un tercer dataset que es la fusión de estos dos últimos. Aquí los resultados:

Dataset fusión “corridos y regional”. En este dataset se analizaron 5141 letras de canciones, los resultados de los términos más comunes de todo el listado de las canciones se muestran en la **Fig. 1**.
Distribución de términos y repetición de entidades:

- El término “te extraño” destaca significativamente, esto sugiere una presencia importante de temas románticos. Varios términos hacen referencia a lugares como: “las vegas”, “nueva york”, “Hong kong” y “estados unidos”. En corridos y música regional es común aludir a ciudades de EE.UU. donde hay comunidades mexicanas o bien actividades delictivas y de negocios.
- Aparecen nombres propios ligados a la cultura del narcotráfico, por ejemplo: “chapo guzmán”, “juan colorado”.

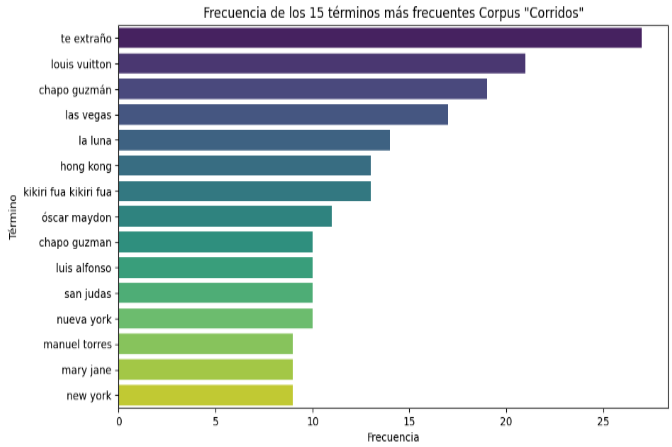

Fig. 2. Los 15 términos más frecuentes en el dataset "Corridos".

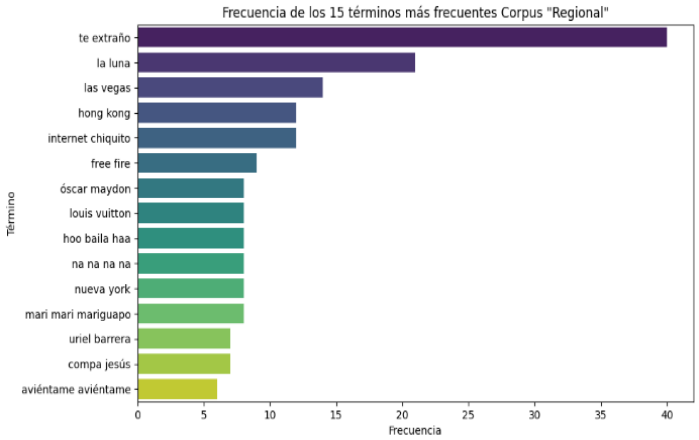

Fig. 3. Los 15 términos más frecuentes en el dataset "Regional".

Dataset "Corridos". En este dataset se analizaron 2,496 letras de canciones, los resultados de los términos más comunes de todo el listado de las canciones se muestran en la **Fig. 2**.

Distribución de términos y repetición de entidades:

- Al igual que en la Fig. 1, el término "te extraño" vuelve a aparecer como el termino con mayor frecuencia. Esto refuerza la idea de que muchos corridos modernos también hay una presencia de temas románticos, más allá de los tradicionales temas de narcotráfico, migración o aventuras.
- El término "louis vuitton" ocupa el segundo lugar. En corridos actuales, es frecuente la mención de marcas de lujo para enfatizar con el poder adquisitivo o el estatus del personaje.

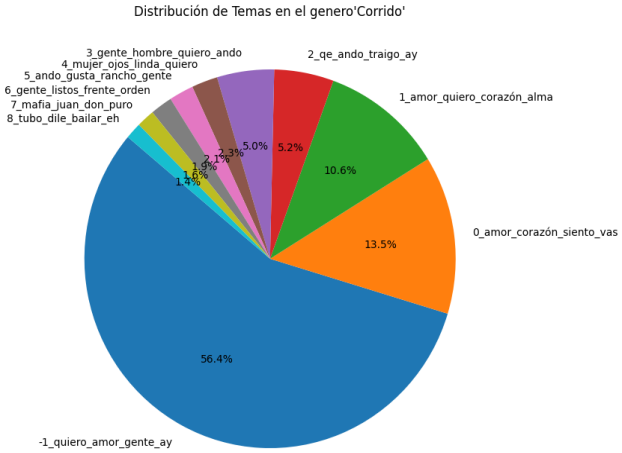

Fig. 4. Distribución de los 10 temas principales del dataset del género 'Corrido'.

Dataset "Regional". En este dataset se analizaron 2,944 letras de canciones, los resultados de los términos más comunes de todo el listado de las canciones se muestran en la **Fig. 3**.

Distribución de términos y repetición de entidades:

- Nuevamente "te extraño" es el término más frecuente, lo que sugiere que la música regional mexicana lo temas de amor y desamor son muy comunes.

- Se menciona "las vegas", "hong kong" y "nueva york", lo que sugiere la influencia de la migración o la conexión con la vida en el extranjero.

b. Modelado de tópicos con BERTopic

Se realizó el análisis temático utilizando el modelo BERTopic, basado en técnicas avanzadas de embeddings y clustering semántico. Los resultados del análisis mostraron diferencias notables en la distribución de tópicos entre los géneros musicales analizados. En el género "Corrido", destacaron tópicos relacionados con emociones personales y afectivas, como "amor_corazón_siento_vas" (13.5%) y "amor_quiero_corazón_alma" (10.6%). Un tema particularmente relevante fue "mafia_juan_don_puro", este tema sugiere que diversas canciones del dataset comparten un contexto común o una combinación temática en la que dichos términos son frecuentes y distintivos. Sin embargo, esto no implica necesariamente que se refieran a un único personaje llamado Juan. Más bien, podría indicar que términos como 'don' (empleado como título respetuoso o jerárquico), 'mafia' (asociado al crimen organizado o contextos ilícitos) y nombres propios como 'Juan', aparecen regularmente juntos en letras que posiblemente reflejan narrativas comunes en los corridos, relacionadas con figuras de poder o historias del narcotráfico. (véase la **Fig. 4**).

Por otro lado, en el género "Regional", los tópicos más frecuentes fueron "amor_quiero_vida_corazón" (33.9%) y "tatara_compa_tatara" (6.2%), resaltando una predominancia en temas relacionados con aspectos emocionales, la vida cotidiana y las

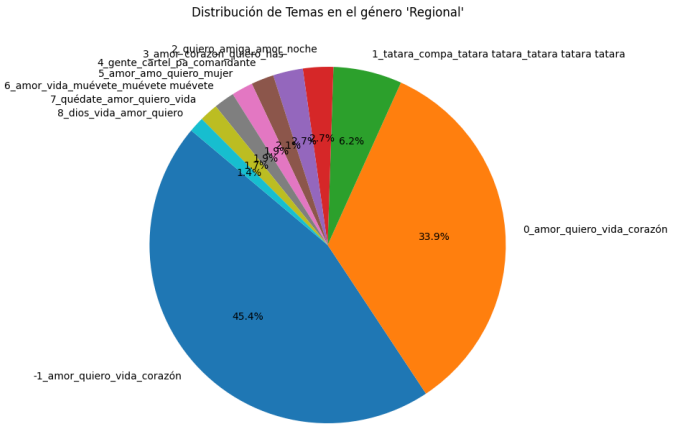

Fig. 5. Distribución de los 10 temas principales del dataset del género ‘Regional’.

Tabla 2. Cantidad de sentimientos en las letras de los corridos y la música regional.

Etiqueta	Corridos	%	Regional	%
Negativo	440	17.63%	1227	41.68%
Neutral	2056	82.37%	1717	58.32%
Positivo	0	0%	0	0%

relaciones personales (véase **Fig. 5.**). Este análisis proporciona una visión clara sobre las temáticas predominantes y evidencia tanto diferencias como similitudes en las narrativas entre ambos géneros.

c. Análisis de sentimientos

El modelo **edumunozsala/roberta_bne_sentiment_analysis_es** fue entrenado utilizando Amazon SageMaker y el contenedor de Deep Learning de Hugging Face. En términos de rendimiento, el modelo alcanzó una precisión del 91.07%, una puntuación F1 de 90.91%, una precisión (*precision*) de 90.64% y una cobertura (*recall*) de 91.18% en el conjunto de datos de prueba. Este modelo es adecuado para tareas de análisis de sentimientos en textos en español, fue aplicado directamente a nuestro dataset ya que el modelo proporciona indicadores útiles sobre la distribución de sentimientos en las letras, estos deben interpretarse como aproximaciones del modelo.

Aunque no existen estudios previos que apliquen exactamente esta combinación de técnicas a los corridos, nuestra metodología se inspira en trabajos que han aplicado modelos similares a otros dominios. Por ejemplo, [9] utilizó RoBERTa para análisis de sentimientos en letras de canciones, mientras que [10] demostró la efectividad de BART-large-MNLI para clasificación de intención en diversos tipos de texto. Nuestra contribución radica en adaptar estos enfoques al dominio específico de la música regional mexicana.

Tabla 3. Cantidad de emociones en las letras de los corridos y la música regional.

Etiqueta	Corridos	%	Regional	%
Tristeza	1048	41.99%	1236	41.98%
Miedo	474	18.99%	559	18.99%
Enojo	300	12.02%	353	11.99%
Alegría	349	13.98%	412	13.99%
Sorpresa	200	8.01%	236	8.02%
Amor	125	5.01%	147	4.99%

Véase en la **Tabla 2** como se representa la frecuencia de ciertos sentimientos en las letras de estas canciones que proporcionan una visión general de cómo se expresan los sentimientos en ambos géneros de música.

d. Análisis de emociones

El modelo **bhadresh-savani/distilbert-base-uncased-emotion**, el cual es una versión de DistilBERT, alcanzó una precisión de 93.8% y una puntuación F1 de 93.79% en el conjunto de datos de emociones de Twitter.

En la **Tabla 3** se representa la frecuencia de ciertas emociones en las letras de estas canciones. Se enfoca en la combinación de ambos géneros (corridos y regional), por lo que proporciona una visión general de cómo se expresan las emociones en ambos géneros de música.

El análisis de 2,944 canciones de música regional y 2,496 corridos revela una fuerte presencia de emociones negativas, especialmente tristeza (42%), miedo (19%) y enojo (12%), reflejando temáticas de dolor, peligro y protesta. Las emociones positivas, como alegría (14%) y amor (5%), están menos representadas, y la sorpresa (8%) aparece ligada a giros narrativos.

En cuanto a sentimientos, predomina un tono neutral: en la música regional con 58%, y en los corridos con 82%, lo que sugiere un estilo más descriptivo o narrativo. Los sentimientos negativos son más notables en la música regional (42%) que en los corridos (18%), y los positivos son casi nulos (0%) en ambos géneros.

e. Detección de intención

Utilizando el modelo BART-large-MNLI y colocando las 3 etiquetas: Glorificar, Describir y Desprestigiar al narcotráfico se realizó el análisis de todas las canciones del repositorio dividido entre las canciones que se catalogaron en corridos y música regional. De esta forma el modelo puede indicar cual es la intención de la letra al menos con el enfoque semántico y contextual que ofrece la inferencia textual del modelo. Después del procesamiento del modelo con todas las canciones se obtuvieron los siguientes resultados:

Como se muestra en la **Tabla 4**, los porcentajes presentados corresponden al promedio de confianza en la consistencia de los resultados, calculado por el propio

Tabla 4. Cantidad de veces que se detectó cada etiqueta de la intención de la canción dentro de cada dataset.

Etiqueta	Corridos	%Conf.	Regional	%Conf.
Glorificar el narcotráfico	81	39%	93	39%
Describir el narcotráfico	1900	45%	2180	44%
Desprestigiar el narcotráfico	515	42%	671	42%

modelo. Aunque estos valores son relativamente bajos, esto se debe a que el modelo consideró las canciones en su totalidad para realizar la inferencia.

También hay que tomar en cuenta otros factores que pueden afectar este análisis general de las canciones. Porque una canción puede solo narrar (describir) cosas referentes al narcotráfico, pero hacerlo desde una perspectiva de protagonista lo que puede llevar a que se describan historias, aspectos o situaciones sobre el narcotráfico de forma positiva no directamente, afectando el análisis por parte del modelo.

5. Conclusiones y trabajo futuro

El análisis de los corridos y la música regional mexicana ha permitido identificar patrones lingüísticos, temáticos, emocionales e intención que reflejan la evolución del género y su impacto en la sociedad. A través del **Reconocimiento de Entidades Nombradas (NER)** y el modelado de tópicos con **BERTopic**, se han detectado referencias recurrentes a emociones personales, lugares internacionales y figuras del crimen organizado, evidenciando la coexistencia de temas afectivos con narrativas sobre narcotráfico y globalización.

El análisis de sentimientos muestra un tono predominantemente negativo, con una ausencia de elementos festivos o positivos, lo que refuerza la idea de los corridos como crónicas de violencia, sacrificio y resistencia cultural. Además, la clasificación de intención con **BART-large-MNLI** resalta la necesidad de diferenciar entre canciones que glorifican el crimen y aquellas que lo narran de manera neutral.

Como líneas futuras de investigación, se propone extender el estudio al análisis transversal del lenguaje en este género musical, explorando cómo han evolucionado las expresiones, temas y referentes culturales a lo largo del tiempo, evidenciando la adaptación de estos géneros a las transformaciones sociales. Además, con el objetivo de mejorar la precisión en las inferencias, se plantea la implementación de un etiquetado a nivel de renglón. Esta aproximación permitiría promediar los valores obtenidos en cada segmento, brindando así un análisis más detallado y fino sobre el contenido y la carga emocional de las canciones.

Agradecimientos. Se agradece el apoyo del Tecnológico Nacional de México campus CENIDET y a la SECIHTI a través de los registros de becas (CVU 1343321, CVU 1105435, CVU 1315217)

Divulgación de intereses. El autor José Eder Martínez Martínez ha recibido apoyo financiero mediante las becas CVU 1343321. El autor Yessenia Díaz Álvarez ha recibido apoyo financiero mediante las becas CVU 1105435. El autor Tlaloc Humberto Vergara Morales ha recibido apoyo financiero mediante las becas CVU 1315217. Los autores Víctor Manuel Millán Aguilar, Gabriel

González Serna y Noé Alejandro Castro Sánchez declaran no tener conflictos de intereses ni apoyos externos relacionados con esta publicación.

Referencias

1. Zhang, Y., Wang, X., Liu, Z.: Advances in Machine Learning for Natural Language Processing. In: J. Artif. Intell. Res., 45(2), pp. 123–145 (2023) doi: doi.org/10.1145/3604931.
2. Brants, T., Franz, A.: Web 1T 5-gram Corpus Version 1.1. In: Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP 2011), Association for Computational Linguistics, pp. 123–130 (2011) url: https://aclanthology.org/D11-1141.pdf
3. Bojar, O., Chatterjee, R., Federmann, C.: Findings of the 2016 Conference on Machine Translation (WMT16). In: Proc. First Conf. Machine Translation, Association for Computational Linguistics, pp. 131–198 (2016) url: https://aclanthology.org/S16-1002.pdf
4. Fell, M., Cabrio, E., Tikat, M.: The Wasabi Song Corpus and Knowledge Graph for Music Lyrics Analysis. Lang. Resour. Eval, 57(1), pp. 89–119 (2023).
5. Kleedorfer, F., Knees, P., Pohle, T.: Oh oh oh whoah! Towards Automatic Topic Detection in Song Lyrics. In: Proc. ISMIR 2008, pp. 287–292 (2008).
6. Wickham, E.N.: Girlbosses, the Red Pill, and the Anomie and Fatale of Gender Online: Analyzing Posts from r/suicidewatch on Reddit. In: Proc. 3rd Int. Conf. Natural Language Processing for Digital Humanities and 8th Int. Workshop on Computational Linguistics for Uralic Languages, pp. 195–212 (2023)
7. Díaz-Santana, M.: Impacto sociocultural de los narcocorridos. Rev. Estud. Mexicanos, 22(3), pp. 34–56 (2014).
8. Serrano, A.: Los narcocorridos: Entre la apología del delito y la crítica social. Rev. Mex. Sociol., 71(1), pp. 51–74 (2009).
9. González-Ruiz, J.: Análisis de sentimiento en letras de canciones usando procesamiento de lenguaje natural. Acta Hispánica, 10(2), pp. 123–140 (2020).
10. Biswas, R.: Flask API for Hugging Face Facebook BART-large-MNLI. GitHub (2021) url: https://github.com/rajatdiptabiswas/flask-api-hugging-face-fb-bart-large-mnli
11. Dataloop AI: BART-Large-MNLI-Yahoo-Answers Model Overview. Dataloop AI (2023) url: https://dataloop.ai/library/model/joeddav_bart-large-mnli-yahoo-answers
12. Stack Overflow: How Does Hugging Face's Zero-Shot Classification Work in A Production Web App? Stack Overflow (2023) url: https://stackoverflow.com/questions/75866093
13. Grootendorst, M.: Bertopic: Neural Topic Modeling with A Class-Based TF-IDF Procedure. (2022) arXiv:2203.05794.
14. Biblioteca Nacional de España (BNE): Corpus de Referencia del Español Actual (CREA). url:https://www.bne.es/es/Catalogos/Corpus
15. Chollet, F.: Keras. (2015) url: https://keras.io