Facultad de Ciencias Económicas y Empresariales

Detección automática de contenido sexual explícito en canciones con modelos largos de lenguaje

Dolores Zamácola Sánchez de Lamadrid
Tutor: Eduardo César Garrido Merchán

E3-Analytics
Clave: 202001782

ÍNDICE DE CONTENIDOS

ÍNDICE DE ECUACIONES

ÍNDICE DE FIGURAS

ABSTRACT

El aumento del contenido sexual explícito en las canciones que escuchan los jóvenes, se ha desarrollado ha sido objeto de inquietud social especialmente por el consumo masivo entre los más jóvenes. El presente Trabajo Fin de Grado propone una solución tecnológica basada en un sistema automatizado de detección de contenido explícito en las letras de canciones. Para poder llevarlo a cabo se ha utilizado un modelo GPT, refinado mediante técnicas de aprendizaje supervisado sobre un corpus de 100 canciones en castellano que han sido previamente etiquetadas por una experta. De esta manera, el modelo será capaz de clasificar canciones por "explícitas" o "no explícitas" con resultados empíricamente sólidos que validan las hipótesis planteadas. De esta manera se plantea como el *Business Analytics* puede contribuir a crear soluciones socialmente responsables que promuevan un consumo musical más consciente y seguro.

1. INTRODUCCIÓN Y MOTIVACIÓN

Nos encontramos en una sociedad en la que prácticamente cualquier tipo de música está al alcance de prácticamente todo el mundo, con independencia de la edad, por ello, analizar el contenido de ésta se ha vuelto un tema de creciente importancia. El presente Trabajo de Fin de Grado se titula *"Detección automática de contenido sexual explícito en canciones con modelos largos de lenguaje*" y tiene como objetivo abordar una problemática en el ámbito musical: la necesidad de identificar y etiquetar las canciones que contienen contenido sexual o violento explícito, con el fin de proteger a los públicos más vulnerables, como los niños y adolescentes. Además, es una realidad que las plataformas musicales cada vez ofrecen más contenido y con ello las letras de sus canciones son cada vez más explícitas, por ello, se ha subrayado la urgencia de aplicar métodos automáticos de detección más eficaces para abordar esta problemática (Chen et al., 2023).

La música siempre se ha considerado una de las manifestaciones artísticas más universales, siendo capaz de generar sentimientos, modificar estados emocionales o incluso tener un impacto en la conducta del que escucha. No obstante, la música no sólo puede hacernos bailar, llorar o cantar sino que también tiene la capacidad de comunicar mensajes tanto implícitos como explícitos, lo que puede llegar a generar impactos no solo beneficiosos sino también perjudiciales en el subconsciente, en nuestra percepción del mundo y en las interacciones humanas.

En los últimos años, se han vuelto muy populares géneros musicales como el reguetón, cuyas letras incluyen contenido que ensalza la violencia, la hipersexualización y el machismo. Prueba de cómo estos géneros han ganado popularidad, es que Bad Bunny – uno de los cantantes más conocidos de reguetón–, fue el artista más escuchado de toda la plataforma de Spotify durante tres años consecutivos (2020, 2021 y 2022) (Spotify, 2022). Conforme este género ha ido ganando popularidad, especialmente entre los más jóvenes, también ha crecido la preocupación acerca de cómo estas impactan en sus oyentes. Grandes psicólogos como Ana Simó reiteran que no solo las melodías sino las letras de las canciones que escuchamos tienen un impacto en la conducta humana, y más aún en los jóvenes, pues se encuentran en un momento de forjar su identidad siendo así más propensos a asimilar los mensajes que reciben (Diario Libre, 2021).

No es ninguna novedad que la conducta sexual y la violencia de género son unos de los grandes problemas sociales de la actualidad , pero investigaciones como las llevadas a cabo por el Instituto de Machos a Hombres (DMAH) demuestran que la constante exposición a canciones con contenido sexual o violento puede provocar efectos en el crecimiento emocional o cognitivo de los adolescentes, especialmente en lo que respecta a la asimilación de valores y patrones de conducta (Milenio, 2020). Es por ello, que la música con contenido explícito tiene un papel importante en la normalización de determinadas conductas, teniendo el reguetón un gran impacto en las visiones de los jóvenes acerca del sexo, las relaciones amorosas y la autoridad (Martino et al., 2006).

Como uno de los géneros más oídos por adolescentes y jóvenes, el reguetón se encuentra en el núcleo de este debate. Un estudio de 2020 analizó 64 temas de este estilo musical y mostró que este género suele incluir menciones explícitas en las que denigran a mujeres o las rebajaban a un objeto reproduciendo estereotipos muy machistas (Díez-Gutiérrez & Muñiz-Cortijo, 2023). En este escenario, es imprescindible preguntarnos si la música que escuchan los jóvenes puede estar influyendo en su forma de ser, de comportarse y en su percepción de las relaciones interpersonales. ¿Podría ser que, al ser normalizadas, las letras de reguetón promuevan conductas agresivas o hipersexualizadas en los jóvenes? La realidad es que, a pesar de que a muchos les atrae estas canciones por su ritmo divertido y pegadizo el problema es, que frecuentemente desconocen el mensaje que las letras están comunicando.

Por todo esto, no podemos negar la necesidad de una solución tecnológica que nos advierta de aquellas canciones que incluyen contenido sexual explícito, violento o que ensalce el machismo. Plataformas como Netflix ya proporcionan alertas acerca del contenido que ofrece cada película y serie notificando si lo hubiese el lenguaje ofensivo, las imágenes sexuales o la violencia que puedan contener. Así, dentro del ámbito musical, la industria discográfica (RIAA), desde los años 90, introdujo las etiquetas “Parental Advisory” para advertir sobre el contenido inapropiado, si bien es verdad, su uso es voluntario (RIAA, 2021). Incorporar un sistema parecido en las canciones proporcionará a los oyentes o sus educadores un mayor control sobre lo que se consume.

Este trabajo de fin de grado sugiere la creación de un modelo de lenguaje largo que identifique de manera automática aquellas canciones con letras de contenido sexual explícito o violento. Así, se podría incorporar a plataformas musicales como Spotify o Apple Music,

este lenguaje que etiqueta cada canción alertando sobre su contenido, proporcionando así una versión más segura de la plataforma. Esto además, posibilita a los padres y educadores limitar el acceso a ciertos temas musicales que no sean adecuados para la edad de sus hijos. No solo esto, sino que igual que en Netflix se podrá crear un perfil "Kids" en las que el contenido esté previamente filtrado.

Un sistema automatizado para etiquetar canciones además, permitirá tornar la información más accesible para los usuarios en general, permitiendo a también los más mayores tomar decisiones acerca de las canciones que escuchan. Además, esta tecnología podría ser utilizada en diferentes campos, como la publicidad o la generación de playlists, contribuyendo a prevenir la exposición indebida a contenido explícito

Concluyendo, la música no es meramente un medio de diversión, sino que sus letras pueden tener un impacto en nuestro comportamiento y actitudes, siendo los niños y jóvenes un público especialmente vulnerable por estar en fase de desarrollo. Es una responsabilidad no solo de los padres sino también de las plataformas musicales de regular el contenido que escuchan los más jóvenes. Con el progreso de la inteligencia artificial y los modelos de lenguaje, nos encontramos más próximos a establecer plataformas musicales que garanticen un consumo consciente de música, salvguardandonos no solo a nosotros sino también a las futuras generaciones de los posibles impactos adversos de las canciones con contenido sexual o violento

La estructura del presente trabajo se dividirá de la siguiente manera: para comenzar, se presenta el contexto y motivación del estudio, después se desarrolla el marco teórico en el que se revisan conceptos clave sobre el contenido explícito en la música, el impacto psicosocial de las letras y el uso de modelos de lenguaje. Seguidamente, se analizará la metodología empleada, en la que se incluye la recopilación de datos y el proceso de etiquetado. La siguiente parte explica el diseño del experimento y las decisiones técnicas adoptadas para el entrenamiento del modelo y después se expondrán los resultados obtenidos, su evaluación y validación de hipótesis. Para finalizar, se recogen las conclusiones, limitaciones y propuestas para trabajo futuro.

2. ESTADO DEL ARTE

La sección del estado del arte nos servirá para revisar el estudio reciente de la aplicación de modelos de lenguaje y técnicas de aprendizaje automático para la detección automática de contenido explícito en distintos formatos. Para analizar esto, se estudiarán tres líneas principales de investigación: en primer lugar, la utilización de modelos basados en transformadores para poder analizar las letras de canciones o textos. En segundo lugar, la evolución de técnicas de aprendizaje profundo y aprendizaje por transferencia aplicadas a la detección de contenido sensible bien en texto, bien en imagen. Y por último, la incorporación de enfoques híbridos, contextuales y éticos que enriquecen la capacidad interpretativa y reguladora de los sistemas de detección automatizada.

Los modelos de lenguaje basados en transformadores han supuesto un punto de inflexión en el proceso del lenguaje natural puesto que estos permiten entablar relaciones semánticas complejas a través de una comprensión profunda del texto. Moviéndonos dentro del ámbito musical, el trabajo de Rospoccher [1] es uno de los referentes, en el que analizando un corpus de más de 800.000 letras de canciones y través del modelo BERT, demuestra que estos modelos superan los enfoques tradicionales en las tareas de clasificación de contenido explícito. No obstante, en este trabajo también se advierte del alto coste computacional que a veces requieren estos modelos. Además en estudios como el de Vázquez Benito [2], se confirma la posición de Rospoccher a través de modelos de transformers es capaz de detectar mensajes de odio en español en redes sociales. Si bien no estamos dentro del ámbito musical, los resultados muestran la versatilidad de estos modelos para identificar contenido sensible.

Por otro lado, el aprendizaje profundo y el aprendizaje por transferencia se han consolidado como estrategias fundamentales para poder desarrollar sistemas de detección precisos y eficientes. En esta línea, el trabajo de Sanz Torres [3] aplica modelos pre entrenados a la detección de contenido sexual en imágenes, subrayando la importancia de curar correctamente los conjuntos de entrenamiento. Esta es una técnica que luego se puede transferir y evolucionar al análisis textual tal y como demuestra Molpeceres Barrientos et al. [4] que comparan clasificadores y codificadores semánticos para poder detectar contenido erótico en textos con especial enfáisis en el rendimiento de los modelos híbridos. Además, Addanki y Murphy [5], crean un sistema de moderación textual mediante redes neuronales profundas, llegando a conseguir un alto nivel de precisión en la clasificación de contenido

sexual durante el curso CS230 en la Universidad de Stanford que se centraba en el aprendizaje profundo.

Por otro lado, esta técnica se ha utilizado también en el ámbito de las humanidades, donde aplicaciones como la de Clerice [6], hacen de puente entre la inteligencia artificial y las ramas más tradicionales. Clerice adaptó modelos de aprendizaje profundo para detectar contenido sexual a nivel de oración en textos latinos del primer milenio. El enfoque que le dió combina el análisis semántico con la clasificación jerárquica, utilizando un corpus valioso para el estudio filológico y así demostrando que las tecnologías más avanzadas pueden aplicarse también a contextos académicos no convencionales.

La complejidad semántica de los mensajes explícitos, que no siempre son fáciles de detectar por la importancia de analizar el contexto entero, ha motivado el desarrollo de métodos híbridos que permiten integrar las reglas lingüísticas en redes neuronales para que se pueda captar el contexto y ver si expresiones son explícitas o no. En esta línea, Okulska y Wiśnios [7], crean un sistema que usa la técnica de resoluscion de correferencias para diferenciar entre contenido erótico benigno y dañino. De esta manera, se demuestra la importancia de atender al contexto discursivo para poder detectar el contenido explícito de manera efectiva.

Siguiendo esta lógica, Markov et al. [8] en su trabajo, fueron más allá y diseñaron un sistema holístico que detecta múltiples tipos de contenido no sexual, como el sexual, violento o discriminatorio. Realizan este trabajo a través de taxonomías específicas, aprendizaje activo y estrategias contra el sobreajuste con implicaciones relevantes para la moderación en redes sociales, donde los textos pueden ser muy variados y complejos.

Aunque estos modelos son muy buenos, su eficacia también ha sido objeto de debate, y algunos investigadores han señalado sus limitaciones. Fell et al. [9] comparan distintos modelos de detección de contenido explícito en letras de canciones y revelan que los modelos más avanzados suelen ser más efectivos, en ellos se deja de lado la subjetividad humana, puesto que es un factor difícil de modelar. En esta misma línea, Darroch y Weir [10], subrayan las discrepancias entre las evaluaciones humanas y las herramientas automáticas. Se basan en la complejidad de interpretar el contenido explícito desde una perspectiva meramente algorítmica puesto que es muy complicado captar el doble sentido, el humor, la ironía o el contexto, que son claves para saber si algo es ofensivo o explícito.

Otros autores han adaptado estos sistemas a contextos particularmente sensibles, como Yu y Yin [11] desarrollan un filtro inteligente para entornos educativos basado en CNNs y lógica difusa. Además, dan un paso más, y lo incorporan a GPT-3 para generar alertas contextuales, lo que eleva el nivel de seguridad dentro de los entornos formativos. En el ámnito legal, Gutfeter et al [12] se centran en la detección de material de abuso infantil (CSAM), tema especialmente delicado y sensible. Implementa clasificadores de extremo a extremo–que aprende el proceso por sí solo– y técnicas de visión por computador, que les permiten detectar patrones visuales problemáticos. Con estos trabajos, se demuestra cómo la precisión técnica debe ir acompañada de principios éticos y de interpretabilidad.

Por otro lado, hoy en día muchas conversaciones con contenido problemático (como acoso o abuso), no ocurren en contextos formales sino en chats, redes sociales y apps de mensajería, por tanto, es necesaria la creación de herramientas que sean capaces de adaptarse al lenguaje coloquial y a las interacciones digitales. Así lo reflejan Colmenares-Guillén y Jiménez-Aguilera [13] en su estudio quienes diseñan un corpus para detectar conversaciones con contenido pederasta en mensajes a través de patrones lingüísticos propios del español. Por último, Povedano Álvarez et al. [14] en su trabajo, realizan una revisión completa de diferentes técnicas de estrategias de aprendizaje que se están usando para detectar contenido sensible, donde subrayan los grandes desafíos tanto éticos, como técnicos y legales que plantea la automatización. Además, de manera similar, Bhatti et al. [15] se centran en detectar contenido explícito visual que detecta pornografia o contenido explícito, con especial énfasis en la clasificación ética y la seguridad en entornos digitales y que sea útil en empresas, colegios o instituciones.

En conjunto, todas estas investigaciones y estudios reflejan el avance de las técnicas de procesamiento del lenguaje natural en el ámbito de la detección automática del contenido explícito y sexual, y la importancia de tener en cuenta las consideraciones contextuales y éticas para su aplicación efectiva. Desde enfoques basados en reglas más simples y tradicionales, hasta modelos más complejos como transformadores, las soluciones que se van planteando, son capaces de integrar técnicas avanzadas de aprendizaje automático con mecanismos de contextualización e interpretación para imitar en mayor medida al humano.

Año	Número	Referencia
2014	[10]	**Darroch, K., & Weir, G. R. S. (2014).** *Measuring Sexually Explicit Content in Text Documents*. *Cyberforensics 2014*.
2018	[15]	**Bhatti, A. Q., et al. (2018).** *Explicit content detection system: An approach towards a safe and ethical environment*. *Applied Computational Intelligence and Soft Computing*.
2019	[9]	**Fell, M., et al. (2019).** *Comparing Automated Methods to Detect Explicit Content in Song Lyrics*. *RANLP*.
2020	[4]	**Molpeceres Barrientos, G., et al. (2020).** *Machine Learning Techniques for the Detection of Inappropriate Erotic Content in Text*. *International Journal of Computational Intelligence Systems*.
2022	[1]	**Rospoccher, M. (2022).** *On exploiting transformers for detecting explicit song lyrics*. *Entertainment Computing*.
2022	[5]	**Addanki, S., & Murthy, N. (2022).** *Text content moderation model to detect sexually explicit content*. *CS230: Deep Learning, Stanford University*.
2023	[2]	**Vázquez Benito, A. O. (2023).** *Desarrollo de un sistema de software basado en Transformers para la detección de lenguaje de odio en medios sociales en español*. *Instituto Tecnológico Superior de Teziutlán*.
2023	[3]	**Sanz Torres, Í. (2023).** *Detección de Contenido Sexual mediante Aprendizaje Profundo y Aprendizaje por Transferencia*. *TFG, UCM*.
2023	[11]	**Yu, Y., & Yin, X. (2023).** *A hypersensitive intelligent filter for detecting explicit content in learning environments*. *Journal of Web Engineering*.
2023	[7]	**Okulska, I., & Wiśnios, E. (2023).** *Towards Harmful Erotic Content Detection through Coreference-Driven Contextual Analysis*. *arXiv*.
2023	[8]	**Markov, T., et al. (2023).** *A Holistic Approach to Undesired Content Detection in the Real World*. *AAAI*.

2023	[6]	**Clérice, T. (2023).** *Detecting sexual content at the sentence level in first millennium Latin texts*. *arXiv*.
2023	[13]	**Colmenares-Guillén, L. E., & Jiménez-Aguilera, J. L. (2023).** *Una aproximación para la detección de contenido sexual en conversaciones digitales*. *CienciAmérica*.
2023	[14]	**Povedano Álvarez, D., et al. (2023).** *Learning Strategies for Sensitive Content Detection*. *Electronics*.
2024	[12]	**Gutfeter, W., et al. (2024).** *Detecting sexually explicit content in the context of child sexual abuse materials (CSAM)*. *arXiv*.

3. ALCANCE DEL TRABAJO

Este Trabajo de Fin de Grado se centra en el desarrollo de un sistema automatizado avanzado de lenguaje natural, para la detección automática de contenido relacionado con la violencia sexual en canciones. A continuación, se presentan los objetivos específicos, las hipótesis que guían los experimentos, así como las asunciones y restricciones del desarrollo del proyecto. De esta manera podemos comprender tanto el propósito como los límites metodológicos del trabajo.

El objetivo principal del proyecto es crear un sistema automatizado avanzado de lenguaje natural para poder detectar el contenido explícito de las canciones en castellano. El proyecto implica la recolección y preprocesamiento de un conjunto de datos de letras de canciones, el refinamiento de modelos de lenguaje de última generación, y la evaluación de su precisión en la identificación de temas de violencia sexual.

De esta manera, a través de los transformadores (como GPT), se podrá identificar, clasificar y etiquetar automáticamente las letras con contenido explícito sexual y servirá a los padres y educadores que deseen cuidar y proteger a los públicos más vulnerables.

Más allá de este fin, también se establecen los siguientes objetivos:

1. Que el modelo de lenguaje personalizado, en este caso GPT, sea un mecanismo útil de detección automática del contenido sensible.
2. Que el modelo esté entrenado de tal manera que se adapte al contexto del reguetón y el trap. Estos estilos tienen un lenguaje y estilo muy concreto y particular: palabras coloquiales, jerga urbana, dobles sentidos…Por tanto, el modelo debe ser capaz de entender la manera en la que estos artistas se expresan para poder identificar y etiquetar las canciones como explícitas.
3. El modelo debe cumplir con un mínimo de rendimiento, que se medirá con unas métricas específicas como la precisión o la sensibilidad. Es decir, una vez el modelo esté entrenado, se analizará si es fiable, eficaz, preciso, y si verdaderamente puede servir como herramienta para clasificar canciones en explícitas o no explícitas.
4. Lo óptimo sería que el modelo se pudiera integrar en plataformas digitales de música para poder integrarlo en el mundo real y que no se quede en un mero experimento. En concreto, serviría como una herramienta no solo que etiquete cada canción con un aviso (como hacen Netflix o HBO), sino que también bloquee automáticamente

canciones con contenido inapropiado para un público vulnerable y así pueda servirse como control parental.

En conjunto, todos estos objetivos realmente articulan los componentes del proyecto, de manera que estos objetivos se alinean con el desarrollo técnico con el propósito social de protección y control del consumo musical.

Por otro lado, en cuanto a las hipótesis formuladas en este trabajo, estas buscan guiar los experimentos y justificar las conclusiones que se obtendrán. Partimos de la premisa que éste trabajo se basa en un experimento que busca comprobar si a través de un customizador de GPT, se puede identificar el contenido sexual o explícito en letras de canciones de reguetón y trap. Como bien se comentaba, para que podamos considerar el experimento como exitoso, el modelo, previamente entrenado, deberá ser capaz no solo de reconocer las frases explícitas y sexuales per se, sino también analizar el contexto semántico de fondo, para poder etiquetar una canción como explícita o no de manera correcta.

Por tanto, se formulan las siguientes hipótesis estadísticamente rigurosas:

En primer lugar, podemos establecer como hipótesis nula (H_0) aquella que afirma que los métodos de clasificación de texto tradicionales y simples, son más eficaces para detectar contenido explícito. Estos modelos se refieren a aquellos que están basados en reglas, diccionarios de términos o modelo estadísticos simples (como por ejemplo, Naïve Bayes o SVM). De esta manera, realmente no habrá diferencia entre los resultados obtenidos de un modelo basado en transformadores y de aquellos de los modelos más tradicionales. Podemos matematizar la hipótesis nula de la siguiente manera:

$$H_0 : \mu_t - \mu_c \leq 0$$

Ecuación 1: Hipótesis nula (H_0)

Donde:

- μt\mu_tμt representa la precisión media del modelo basado en transformadores (GPT personalizado).
- μc\mu_cμc representa la precisión media del modelo clásico (basado en técnicas tradicionales).

Por otro lado, la **hipótesis alternativa (H_1),** establecería que los resultados que obtengamos tras realizar los experimentos con los transformadores, por su capacidad para interpretar secuencias lingüísticas amplias, serán más precisos, con mayor capacidad para generalizar, con un equilibrio más sólido. En este caso, la matematizamos de la siguiente manera:

$$H_1 : \mu_t - \mu_c > 0$$

Ecuación 2: Hipótesis alternativa (H_1)

Donde:

- μt\mu_tμt representa la precisión media del modelo basado en transformadores (GPT personalizado).
- μc\mu_cμc representa la precisión media del modelo clásico (basado en técnicas tradicionales).

Para poder llevar a cabo la validación de estas hipótesis se hará a través de una serie de pruebas comprobatorias que evaluarán el rendimiento del modelo en cada una de las fases. Así, se podrá comparar el rendimiento entre el nuevo modelo personalizado frente a otro más tradicional. Todo esto se hará mediante análisis estadísticos para poder analizar de manera objetiva las diferencias de ambos modelos y demostrar que no son producto del azar.

Por otro lado, para poder llevar a cabo este Trabajo de Fin de Grado, se han tenido en cuenta una serie de asunciones que han permitido establecer un punto de partida estable sobre el cual se ha construido el diseño, entrenamiento y validación del modelo.

1. Para comenzar, se ha asumido que las canciones seleccionadas para entrenar el modelo son una buena muestra que refleja el estilo y el lenguaje del reguetón y trap y su espectro explícito y no explícito. Esto es lo que llamamos generalización espacial, que supone que con la muestra de varias canciones, el modelo podrá aplicarse a otras canciones que tengan un estilo similar.
2. Por otro lado, otra asunción importante es la generalización temporal, es decir, que los patrones que sigue la muestra, se mantienen, más o menos estables a lo largo del tiempo. Es decir, confiamos que el tipo de lenguaje, expresiones, metáforas o jergas

que se utilizan a día de hoy en estos géneros, no cambiará demasiado rápido con el tiempo. Esto permitirá que el modelo sea útil, no se desactualice y no haya que estar entrenándolo constantemente.

3. Además, se asume que el modelo, una vez entrenado, actuará de manera objetiva y sin incorporar sesgos adicionales durante la predicción. Esto es importante porque aunque partas de datos balanceados, los propios algoritmos pueden amplificar patrones problemáticos si no se supervisa adecuadamente.
4. Asimismo, asumimos que el etiquetado manual realizado por la experta durante el entrenamiento es preciso y coherente. Decidir qué expresiones son explícitas o no, es una tarea subjetiva, por tanto es crucial que la persona que lo etiqueta lo haga de manera uniforme y aplicando los mismos criterios apra todas las canciones. Si esto no fuese así, el modelo puede aprender a replicar sesgos o errores de etiquetado y el modelo estará por tanto, mal entrenado.
5. Por último, se asume que los datos y canciones usadas para el entrenamiento del modelo, son buenos, variados y equilibrados para poder evitar sesgos importantes. Para garantizar esto, se han utilizado 100 canciones para el entrenamiento: 50 explícitas y 50 no explícitas, habiendo canciones de distintos artistas y estilos para que el modelo esté predispuesto a ciertos patrones.

Estas asunciones representan el marco conceptual sobre el que se construye el experimento, y permiten interpretar los resultados dentro de un contexto realista y bien fundado.

Por otro lado, para poder llevar a cabo el presente trabajo de fin de grado de manera exitosa, se enfrentarán una serie de restricciones que podrán influir a la hora de conseguir unos resultados óptimos. Estas limitaciones de carácter técnico, temporal, metodológico y contextual se deben tener en cuenta a la hora de valorar tanto la aplicabilidad como la generalización del sistema que se ha propuesto.

1. Para comenzar, la primera restricción es relativa a la capacidad computacional, puesto que para entrenar modelos basados en transformadores desde cero, se requiere una considerable cantidad de recursos hardware para poder entrenarlos y ajustarlos. Debido a esta falta de recursos, aprovecharemos un modelo que ya viene entrenado (transfer learning), como GPT, y lo personalizaremos para ajustarlo al caso.
2. Por otro lado, nos encontramos con un limitado tiempo para el desarrollo del trabajo de fin de grado, que a su vez, debe compatibilizarse con la continuación de los

estudios. A la hora de explorar modelos más complejos o de realizar hipótesis o experimentos adicionales, esto será una de las grandes restricciones.

3. Además, a la hora de obtener datos y etiquetarlos, nos encontraremos con restricciones significativas. Esto se debe a que la elección de las canciones seleccionadas para el entrenamiento y su clasificación como explícitas o no explícitas se deberá hacer manualmente, lo que implica gran subjetividad por parte del experto.
4. Además, se encontraran complicaciones para encontrar canciones de reguetón y trap que no contengan contenido explícito, sexual, o violento, puesto que la manera de expresarse de estos artistas suele ser provocadora directa y cargada de metáforas sexuales. Esta característica complicaría la búsqueda equitativa de canciones consideradas "no explícitas", ya que incluso aquellas que no contienen frases o palabras explícitas de forma evidente, pueden ocultar connotaciones implícitas difíciles de detectar.
5. También habrá que tener en cuenta las restricciones éticas y legales puesto que la mayoría de las canciones tienen derechos de autor. Como este trabajo tiene su base en un corpus compuesto por canciones de trap y reguetón que pertenecen a artistas reconocidos y están protegidas legalmente.
6. Por último, existe una limitación de géneros musicales:únicamente nos centraremos en canciones en castellano en trap y reguetón, en su manera de expresarse, metáforas… por tanto, se dejarán fuera muchos otros estilos que, aunque serían interesantes de analizar, como no utilizan el mismo lenguaje, la efectividad del modelo podría no ser extrapolable a otros estilos.

Todas estas limitaciones claramente delimitan el desarrollo y el alcance realista del proyecto, permitiendo evaluar los resultados obtenidos dentro de un marco teórico y ético bien fundado.

En definitiva, este apartado ha permitido establecer los pilares fundamentales que sustentan este trabajo: empezando desde unos objetivos claramente definidos, hasta unas hipótesis rigurosamente formuladas, que permitirán contrastar el rendimiento del modelo propuesto frente a las alternativas más simples y tradicionales. Además, se han expuesto una serie de asunciones razonables que permiten dar solidez metodológica al experimento, junto con las restricciones que enmarcan los límites reales de la investigación. Todo ello orienta el diseño del modelo y los experimentos que se realizarán, para poder garantizar que los resultados sean interpretados de manera crítica y precisa, en coherencia con las condiciones bajo las que

se han obtenido. De esta manera, se ha delimitado el alcance del trabajo, lo que servirá de base para el desarrollo de las siguientes fases del estudio.

4. MARCO TEÓRICO

a. Descripción cualitativa del problema

Las canciones con contenido explícito y sexual, violento e hipersexualizado, especialmente en canciones de reguetón y trap, están a la orden del día, lo que ha sido objeto de preocupación por parte de padres, educadores y psicológicos. La música ha sido una herramienta de expresión social y artística desde hace muchos años, pero el problema viene cuando géneros musicales normalizan discursos que cosifican el cuerpo y hablan de manera muy directa sobre sexo. Esta preocupación se incrementa cuando los que escuchan estos discursos son niños o adolescentes puesto que al estar en la fase de desarrollo, son altamente influenciables, y pueden llegar a asimilar sin filtro estos contenidos.

En este sentido, la problemática que aborda este trabajo no es solo una cuestión de gusto musical o de libertad de expresión, sino que va más allá: la profunda preocupación hacia la pedagogía de los más jóvenes, que asimilan valores y educación sin darse ni cuenta.

No obstante, uno de los grandes desafíos a la hora de abordar este fenómeno, es que no siempre está tan claro qué podemos considerar "contenido explícito" y que no. ¿Cuál es el límite? ¿Quién decide qué es apropiado? ¿Cómo distinguimos lo romántico o erótico y lo sexualmente violento? Como las respuestas a todas estas preguntas son ambiguas, la creación de un modelo que detecte de manera automática las letras explícito sexuales no es una tarea fácil. No solo habrá que detectar las palabras o expresiones que "crucen la raya", sino que habrá que dar un paso más: entender el contexto, dobles sentidos, o incluso las expresiones culturales que cambian según el país. Además, estos géneros musicales utilizan un lenguaje muy específico con jerga urbana y expresiones como "bellaquear", "perreo", "mamacita", que no son palabras recogidas en diccionarios pero que pueden a su vez tener una fuerte carga sexual. Por tanto, el modelo debe adaptarse a los usos lingüísticos de estos géneros y no solo a reglas lingüísticas rígidas y tradicionales.

Parece que la sociedad está respondiendo a esta preocupación, y como bien se comentaba, redes sociales o plataformas de streaming ya han incorporado un control para regular la difusión del contenido explícito. El sector musical no se puede quedar atrás y por ello,

también debe tomar medidas para dar una respuesta socialmente responsable frente a esta necesidad emergente de regular las canciones que escuchamos.

A modo de esquema visual, podríamos representar el problema con la siguiente figura:

Figura 1. Problemática de la detección del contenido explícito

b. Metodología de un modelo grande de lenguaje

Antes de abordar la metodología basada en modelos de lenguaje grandes, es importante recordar cómo han funcionado hasta ahora los enfoques más tradicionales que han sido utilizados para detectar contenido explícito. Dentro de los métodos más sencillos, destacan aquellos que aplican diccionarios de palabras prohibidas o listas negras, cuyo rendimiento suele ser limitado. Así un estudio encontró que el filtrado por diccionario logra apenas un 61% de F1-score, evidenciando su incapacidad para captar el contexto semántico (Chin et al., 2018).

A modo de contraste, aquellos basados en aprendizaje automático suelen tener ser más eficientes y mostrar un rendimiento superior. A modo de ejemplo, utilizando técnicas de bolsa de palabras y árboles de decisión, Chin et al. (2018) lograron un 78% de F1-score en la detección de letras explícitas en la lengua coreana, lo que demuestra cómo estos modelos superan claramente al método basado en diccionario. Estos resultados refuerzan la necesidad de enfoques más sofisticados, capaces de interpretar no sólo las palabras individuales, sino también su significado dentro del contexto.

Dentro de los programas más avanzados en el campo de la inteligencia artificial, destacan los modelos grandes de lenguaje, también conocido como LLMs. En los últimos años estos programas se han convertido en herramientas fundamentales dentro del ámbito del *Business Analytics* y podríamos definirlos como un modelo de *deep learning* entrenado con grandes cantidades de texto. Estos utilizan unas arquitecturas de tipo transformer contando con una amplia capacidad para entender, generar y clasificar textos en lenguaje natural casi como si fueran personas. GPT (*Generative Pre-trained Transformer*) es uno de los modelos de tipo transformer, familia de modelos desarrollados por OpenAI, y ha demostrado que puede llevar a cabo tareas de automatización, de análisis de sentimiento y de clasificación semántica.

El equipo de Vaswani fue el primero en introducir la tecnología transformer en el año 2017, dando el gran paso hacia la sustitución de métodos más tradicionales como RNN o LSTM (Vaswani et. al., 2017). La gran novedad que trajeron consigo los transformers es su capacidad para procesar todo un texto de una vez sin tener que ir analizando palabra por palabra puesto que utilizan mecanismos de atención (*self attention*). Estas técnicas permiten identificar y conectar palabras relacionadas entre sí aunque estén muy separadas, lo que es clave a la hora de analizar canciones puesto que el significado de una palabra o frase varía en función del contexto. En contextos de *Business Analytics*, la capacidad de los LLM como GPT para comprender matices del lenguaje y clasificarlos, resulta muy valiosa porque lo capacita para automatizar el análisis de textos y adoptar decisiones basadas en estos análisis de forma rápida y escalable.

A continuación, se expone un esquema visual de las fases por las que pasa la metodología de un modelo grande de lenguaje:

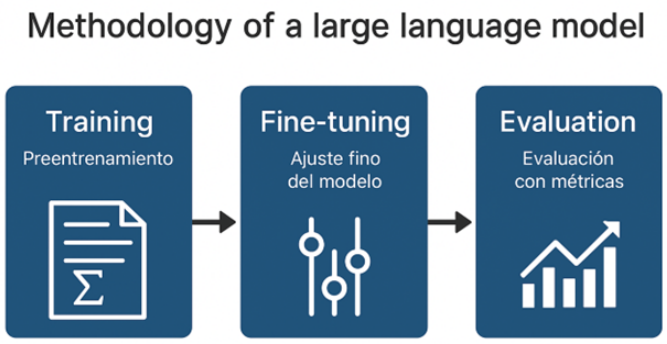

Figura 2. Fases del desarrollo de un modelo grande de lenguaje (LLM)

Para comenzar, en términos metodológicos, los modelos de tipo transformador como GPT primero se entrenan con una gran cantidad de texto, lo que denominamos, *pre-training,* y que servirá de base sólida para luego construir una tarea concreta. En esta fase del entrenamiento, el modelo aprende y adquiere un profundo entendimiento del lenguaje para que pueda entender la manera en la que se estructuran las frases, qué palabras suelen ir juntas… Aquí es donde el modelo aprende a desarrollar una representación estadística del lenguaje, por lo que ya podrá predecir qué palabras vienen después de la oración. La manera en la que el *pre-training* funciona, se puede explicar a través del entendimiento de su arquitectura:

La base del transformer y de GPT es el mecanismo de atención (*attention*), que asigna pesos a cada palabra de acuerdo con su relevancia con respecto a los demás en una secuencia (Vaswani et al., 2017). El mecanismo de *attention* se formaliza mediante las matrices de *Consulta* (Q), *Clave* (K) y *Valor* (V), obtenidas a partir de los vectores de entrada del modelo. Podríamos decir que la consulta, representa lo que estamos buscando, la clave, las palabras disponibles a las que se puede prestar atención y el valor, la información real que se va a usar si se decide prestar atención a esa palabra.

Esto quiere decir que con este mecanismo, a cada palabra (o token) del texto se le proyecta a estas tres representaciones mediante matrices de pesos que van aprendiendo (W^Q, W^K, W^V). Esto quiere decir que el modelo aprende matemáticamente cómo transformar cada palabra en Q, K o V usando multiplicaciones con matrices que él mismo va ajustando durante el entrenamiento. La *atención* que mostrará el modelo entre una posición i y una posición j del texto, la calculará como el producto escalar entre sus representaciones Q_i y K_j, determinando así cuánto *"atiende"* la posición i al contenido de la posición j. Es decir, el modelo compara lo que busca (Q), con lo que cada palabra ofrece (K) usando una fórmula de similitud que le ayudará a ver cuánto "caso" debería hacerle a esa palabra.

No obstante, antes de calcular la ponderación, se escalan estos productos por \sqrt{d_k} (siendo d_k la dimensión de los vectores K) para estabilizar los gradientes durante el entrenamiento. Esto quiere decir que para que los números no se desordenen y el modelo no se sature, se divide todo por la raíz cuadrada de la dimensión del vector, que mantiene la estabilidad del modelo a la hora de aprender.

Finalmente, se aplica una función *softmax* que convierte estos puntajes en probabilidades o porcentajes (valores entre 0 y 1 que suman 1), las cuales se usan para ponderar una

combinación lineal de los valores V, es decir, para saber cuánta atención hay que darle a cada palabra.

Matemáticamente, podríamos expresar el mecanismo de atención que se acaba de explicar de la siguiente manera:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{Q\,K^T}{\sqrt{d_k}}\right) V$$

Ecuación 3: Mecanismo de atención escalar

Donde:

- Q (queries): representa lo que queremos entender (por ejemplo, el significado de una palabra).
- K (keys): representa las palabras con las que se va a comparar.
- V (values): es la información que el modelo va a usar si una palabra resulta relevante.
- $\sqrt{d_k}$: es una constante de escala para que los valores no se disparen.

En esta ecuación, Q,K^T/\sqrt{d_k} realiza una matriz de puntuajes de similitud entre cada par de palabras, es decir, compara cada palabra con todas las demás para ver cuánto se parecen o se relacionan a través del producto escalar. Con este punto de partida, la función *softmax* transforma en las distribuciones de atención para decidir cuán importante es cada palabra para entender el contexto. Al multiplicarlo por V, el resultado que se obtiene es una representación ponderada de las palabras, es decir, el modelo mezcla toda la información de todas las palabras según esas probabilidades de manera que el modelo no solo "ve" la palabra individual, sino una versión mejorada de ella incluyendo lo que se relaciona con las otras. De esta manera, en el resultado, cada posición incorpora información contextual relevante de las otras posiciones. Así, este esquema permite al modelo captar relaciones de largo alcance en el texto, algo que es crucial para detectar contenido explícito puesto que una palabra o frase puede depender de varias palabras en conjunto.

Dicho de manera más simple, este mecanismo es el que permite al modelo crear una representación mejorada de cada palabra, de manera que tiene en cuenta todas las palabras con las que una palabra se relaciona, captando relaciones a largo plazo en el texto. Esta es la

razón por la que el modelo es capaz de detectar contenido explícito ya que este muchas veces depende del contexto completo.

Por otro lado, lo que hace a los transformers un mecanismo tan potente, es que utilizan atención múltiple por cabezas (*multi-head attention*), que quiere decir que es capaz de ejecutar varios mecanismos de atención a la vez y en paralelo (cabezas), siendo capaz de aprender distintos patrones o relaciones. Cada cabeza h_i calcula su propia atención con proyecciones W_i^Q, W_i^K, W_i^V diferentes, y sus salidas se concatenan y combinan linealmente para formar la salida final de la capa de atención:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_H)W^O, \quad \text{donde} \quad \text{head}_i = \text{Atención}(QW_i^Q, KW_i^K, VW_i^V)$$

Ecuación 4: Atención multi-cabeza (Multi-Head Attention)

Donde:

- Q: *Query* (consulta), una representación del elemento actual para el que queremos calcular la atención.
- K: *Key* (clave), una representación de los elementos con los que se compara el elemento actual.
- V: *Value* (valor), la información que se va a combinar en función de la atención calculada.
- W_i^Q, W_i^K, W_i^V : matrices de pesos aprendidos específicas para la cabeza iii, que transforman las matrices de entrada Q,K,V antes de aplicar la atención.
- Attention($\cdot$): la función de atención que calcula los pesos de importancia entre palabras (como en la fórmula escalar anterior).
- $head_i$: resultado de aplicar la atención en la cabeza iii.
- H: número total de cabezas de atención (por ejemplo, 8 o 12)
- Concat($\cdot$): operación que **concatena** (une horizontalmente) las salidas de todas las cabezas.
- W^o: matriz de pesos final que se aplica a la concatenación de las cabezas para generar la salida final.

En esta ecuación de *multi-head attention*, H se refiere al número de cabezas de atención paralelas que ha sido calculada tras calcular la atención por separado en cada una de estas

cabezas y se concatenan todas sus salidas. Así esta concatenación se multiplica por una matriz de pesos final, que es a lo que se refiere W^O. Así, el modelo puede capturar distintos tipos de relaciones semánticas simultáneamente, por ejemplo, mientras una cabeza se centra en relaciones de género/prenombre, otra se focaliza en lenguaje vulgar y así…

Después del bloque de atención, cada capa del Transformer incluye una *red feed-forward,* que simplemente son un par de capas densas (*fully connected)* que transforma cada posición de forma no lineal (como ReLU o GELU). Esto utiliza mecanismos de normalización (*LayerNorm*), que estabiliza los valores numéricos y adición de conexiones residuales, que suman la entrada de cada bloque a su salida, para facilitar el entrenamiento profundo aun cuando no hay muchas capas.

En resumen, podríamos representar visualmente el funcionamiento de la arquitectura de GPT de la siguiente manera:

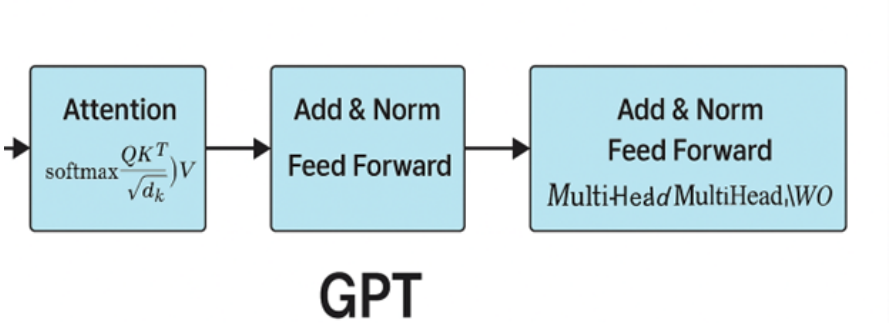

Figura 3. Estructura interna de una capa del modelo GPT

Toda esta arquitectura en conjunto, permite que modelos como el de GPT representen con fidelidad complejas dependencias en texto. A modo de matiz, cabe destacar que los modelos GPT suelen emplear solo la parte decodificadora del transformer, y sólo aplican máscaras causales a las matrices de K y V de manera que cada posición del texto pueda "ver" sólo las palabras anteriores propio del entrenamiento auto-regresivo. Este es el modelo que se utiliza para generar texto puesto que no puede adelantarse a palabras que todavía no han generado. Esto difiere de otros modelos basados en transformers como por ejemplo BERT que utilizan la parte codificadora y atención bidireccional (Devlin et al., 2019), de manera que el modelo puede "ver" tanto hacia delante como hacia atrás. No obstante, para efectos de la tarea de clasificación (detección de contenido explícito) el mecanismo fundamental de atención sigue siendo el mismo, GPT.

Una vez comprendida la arquitectura y como se preentrena el modelo GPT, este se adapta a la tarea mediante *fine-tuning,* comenzando así la segunda fase. Este concepto se refiere a volver a entrenar el modelo mediante el reajuste de los pesos del modelo GPT a través de un conjunto de entrenamiento específico. De esta manera, para que el modelo pueda especializarse en la tarea deseada, se entrena con un conjunto de datos más pequeño, y que ha sido previamente preparado para poder ajustar sus parámetros internos, basándose en los nuevos datos introducidos. Así el modelo detectará patrones específicos de la tarea concreta.

En el proceso de *fine-tuning* se agrega una capa densa al final del modelo (un clasificador) que toma la representación generada por GPT y emite una probabilidad de pertenencia a una clase, como podría ser "contenido sexual explícito". De esta manera, el segundo entrenamiento ajusta los pesos minimizando una función de pérdida apropiada (por ejemplo, la entropía cruzada binaria, que penaliza más las clasificaciones erróneas).

De esta manera, podemos estructurar el modelo GPT a través el siguiente algoritmo de transferencia del aprendizaje el proceso de personalización del modelo GPT puede estructurarse mediante el siguiente algoritmo de transferencia del aprendizaje:

1. Cargar el modelo GPT preentrenado sobre grandes corpus de texto general (lenguaje natural).
2. Añadir una capa final de clasificación binaria que predice si una letra es explícita o no.
3. Cargar el corpus específico de 100 canciones etiquetadas manualmente por la experta.
4. Reentrenar solo las últimas capas del modelo, ajustando los pesos con un *learning rate* bajo para no sobreescribir el conocimiento general.
5. Evaluar el modelo sobre canciones nuevas no vistas para comprobar su rendimiento y ajustar si es necesario

Por tanto, se podría decir que debido al conocimiento del lenguaje que ya tiene GPT (por su preentrenamiento masivo), el modelo converge rápidamente y logra capturar con precisión las sutilezas como podría ser en este caso, del lenguaje soez, coloquial y metafórico presentes en las letras de reguetón y trap. Esto demuestra el gran poder del *transfer learning* en NLP, que no entrena el modelo desde cero, sino que es capaz de especializarse con una pequeña muestra de datos científicos, aprovechando el entrenamiento de un modelo genérico, con el objetivo de lograr resultados sobresalientes en menos tiempo. Es por esto, que los últimos

modelos GPT, como por ejemplo GPT-3, que cuenta con 175.000 millones de parámetros, han demostrado tener capacidades cada vez más sorprendentes para adaptarse y especializarse en tareas muy específicas dentro del lenguaje natural, a través de la detección de patrones sensibles con alta precisión (Brown et al., 2020).

Así, en el presente Trabajo de Fin de Grado, se pone en práctica esta capacidad de adaptarse de los LLM, personalizando un GPT para poder identificar de manera automática contenido explícito sexual en texto.

En tercer y último lugar, entramos en la fase de evaluación a través de métricas. Esta fase es totalmente esencial para poder evaluar de manera objetiva el modelo que ha sido preentrenado y personalizado. Para llevar a cabo esta evaluación se podrá hacer a través de diferentes métricas de clasificación binaria como la precisión, la sensibilidad, la especificidad y la exactitud. A través de estas métricas se conseguirá medir el desempeño tanto antes como después del proceso de *fine-tuning* de GPT para obtener una comparación cuantitativa del valor que añade el ajuste personalizado.

A continuación se definen cada una de ellas en el contexto del trabajo que nos interesa:

- La **precisión (*precision*)** o valor predictivo positivo, se refiere a aquella proporción de predicciones que el modelo ha clasificado de manera correcta, es decir, cuán de precisos son los positivos detectados por el modelo.
- La **sensibilidad (*recall* o *true positive rate*)** mide la capacidad de detectar los verdaderos positivos
- La **especificidad (*true negative rate*)** se refiere a la capacidad de descartar los negativos evitando los falsos positivos.
- La **exactitud (*accuracy*)** es aquella proporción global de aciertos del modelo, refiriéndose tanto a los positivos como los negativos

Estas métricas se calculan a través de la matriz de confusión que es capaz de resumir los aciertos y errores del modelo en cada categoría. En el presente caso, la matriz de confusión será de 2x2 y recogerá los Verdaderos Positivos (VP), Falsos Positivos (FP), Falsos Negativos (FN) y Verdaderos Negativos (VN). Una vez se tienen esos valores ya se pueden derivar las métricas mencionadas: por ejemplo, Precisión = \frac{VP}{VP+FP}, Sensibilidad = \frac{VP}{VP+FN}, etc.

Podemos concluir que, es a través de esta metodología en la que se coge un modelo preentrenado, se realiza un segundo entrenamiento de *fine-tuning* para ajustarlo y con una evaluación final a través de métricas como la precisión, sensibilidad, especificidad y exactitud, se podrá lograr un modelo que cumpla una finalidad específica. Así, siguiendo esta metodología, se entrenará al modelo para que sirva a su propósito principal: detectar las canciones explícito sexuales para poder hacer consciente a los oyentes del contenido de las canciones que escuchan.

5. EXPERIMENTOS

A. Diseño del experimento

Como bien se decía, este trabajo de fin de grado se ha centrado en la creación, ajuste y evaluación de un modelo de lenguaje personalizado y en este apartado se estudiará cómo se ha llevado a la práctica toda la parte metodológica.

Como bien ya se ha dejado claro: el modelo de lenguaje elegido ha sido GPT con un *fine-tuning* para que sea capaz de clasificar letras de canciones en dos categorías: "explícita" o "no explícita". Para ello se han seguido las siguientes fases, que han sido meticulosamente planeadas:

1. Creación y estructuración del corpus

El primer paso en el diseño del experimento ha sido recolectar datos para crear un corpus propio. Preparar el corpus es una tarea esencial puesto que de ello depende que el modelo aprenda de manera efectiva. Así se ha compilado el conjunto de letras de canciones compuesto por 100 canciones en castellano, de reguetón y trap puesto que estos géneros presentan una gran carga de contenido explícito sexual.

Este paso es el primero que se dio en búsqueda de cumplir el objetivo de desarrollar el sistema automatizado basado en lenguaje natural para poder identificar contenido explícito. Además, la creación de un buen corpus será una herramienta totalmente esencial para que el modelo se adapte al lenguaje específico del reguetón y el trap puesto que de esta base es de la que aprenderá.

Para la creación del corpus se ha requerido la limpieza y transformación de las letras de canciones seleccionadas. De entre estas 100 canciones, 50 tenían referencias sexuales de

forma directa, implícita o metafórica y 50 no contenían este tipo de contenido, para que esta fase no estuviera sesgada. Todas las canciones seleccionadas utilizaban un lenguaje común que incluía el uso de la jerga urbana y estructuras narrativas y palabras no convencionales.

La etiquetación según explícitas o no explícitas se llevó a cabo de manera manual por una experta, que, aunque sea una tarea laboriosa y añada un grado de subjetividad, garantiza un análisis profundo que permite entender el contexto de fondo. Se prefirió hacer de esta manera y no con otros sistemas automáticos (como el uso de listas de palabras prohibidas) porque la ambigüedad y riqueza lingüística del reguetón y el trap exigen una interpretación que un enfoque basado en un diccionario no es capaz de ofrecer. Todas las letras de las canciones se recopilaron en un documento estructurado de Word que ha constituido el corpus final.

2. Análisis de frases explícitas y tabla de referencia

Antes de entrenar el modelo se elaboró una tabla específica (Tabla 1) en la que, para las canciones etiquetadas en la base de datos como "explícitas", se recogieron las frases concretas que se consideraban explícitas. Esta tarea fue clave puesto que ayudó al modelo a aprender qué frases o palabras escondían un contenido sexual entrenándolo para detectar no solo palabras explícitas, sino también dobles sentidos, metáforas sexuales, jerga urbana y estructuras normativas del género. Algunas de las frases más representativas eran parecidas a las siguientes:

- "Te voy a dar hasta que Dios diga"
- "Ese booty que hasta un ciego pueda ver"
- "Me mete sudando su cuerpo de mora"

De esta manera, este paso se vincula estrechamente con el segundo objetivo: que el modelo entrenado se adapte al lenguaje propio del reguetón y trap puesto que sin estas referencias, el modelo pasaría por alto muchas expresiones o metáforas que a simple vista pueden no reconocerse como explícitas.

3. Entrenamiento del modelo (*fine-tuning* con GPT)

Como bien se ha dicho en el apartado de la metodología, se ha utilizado una técnica de aprendizaje supervisado (*fine-tuning)* sobre el modelo GPT preetrenado. Se eligió GPT por varias razones: primero, por eficiencia computacional puesto que entrenar un modelo desde

cero hubiese sido demasiado costoso. Y segundo, porque GPT ya cuenta con un profundo conocimiento de la lengua castellana y tercero, por su capacidad contextual.

En términos algorítmicos, el fine-tuning se realizó utilizando un algoritmo de aprendizaje supervisado basado en descenso de gradiente estocástico, optimizado con AdamW, que minimiza una función de pérdida binaria (entropía cruzada). Esta función penaliza las predicciones incorrectas, permitiendo ajustar los pesos del modelo GPT para adaptarlo a la tarea de clasificación binaria (explícito / no explícito).

De manera más visual y simplificado, podemos observar en los siguientes pasos, que simplifican el proceso que sigue el algoritmo durante el proceso simplificado:

1. Inicializar GPT con pesos preentrenados.
2. Añadir capa de clasificación binaria.
3. Cargar dataset etiquetado (explícito / no explícito).
4. Calcular la entropía cruzada entre predicción y etiqueta real.
5. Optimizar pesos usando AdamW (descenso de gradiente).
6. Repetir hasta convergencia o alcanzar número de épocas.

Esta parte del proceso, es clave no solo para la elaboración de un modelo que se centre en la tarea específica de etiquetar canciones, sino que también permitirá entender bien el lenguaje propio del reguetón y el trap.

Además, en esta fase, se vigiló muy de cerca el sobreajuste (*overfitting*), para que el modelo no se aprendiese el modelo de entrenamiento sino que con los ejemplos como base, fuese capaz de generalizar. Por otro lado, una parte del corpus fue reservado como conjunto de validación, es decir, se apartaron una serie de canciones que no fueron introducidas en la base de datos para que más tarde nos sirviese para ver si funcionaba de manera efectiva. Tras haber hecho un primer entrenamiento con la base de datos y haberlo validado con canciones que el modelo no había "visto" antes, se aplicó un sistema de *feedback* y reajuste. Es decir, se estudiaron los errores que tuvo el modelo en este primer entrenamiento- como frases que no había clasificado como explícitas y viceversa- y se corrigieron dándole *feedback* al modelo.

4. Evaluación mediante métricas y análisis

Una vez completadas las fases de preprocesamiento, entrenamiento y *fine tuning* del modelo GPT, llegamos a un pilar fundamental del *Business Analytics:* la evaluación y validación empírica de los resultados obtenidos para poder evaluar el rendimiento del modelo. Esto proporciona evidencias objetivas que permitirán validar o refutar la hipótesis planteada y está directamente alineado con el tercer objetivo propuesto: que el modelo obtenga un mínimo rendimiento.

La evaluación del modelo se hará a través de métricas de clasificación binaria que nos permitirán tener una visión completa del comportamiento del modelo. Como ya ha sido comentado en la metodología, estas son: precisión, sensibilidad, especificidad y exactitud. Estas métricas han resultado ser efectivas cuando estamos hablando de clasificación binaria ("explícitas" y "no explícitas") y se calculan a partir de la matriz de confusión, que resume de manera numérica los aciertos y errores que realiza el modelo en cada categoría. Así, la matriz contempla cuatro posibilidades en la labor de clasificación de las canciones: TP (Verdaderos Positivos o *True Positives*), TN (Verdaderos Negativos o *True Negatives*), FP (Falsos Positivos o *False Positives*) y FN (Falsos Negatives o *False Negatives*).

A partir de los valores que resulten de poner a prueba el modelo y que figuran en la matriz de confusión, se derivarán las métricas que orientan el análisis posterior tras aplicar las siguientes fórmulas para cada métrica:

$$\text{Precisión} = \frac{VP}{VP + FP}, \quad \textit{Ecuación 5: Precisión (Precision)}$$

$$\text{Sensibilidad} = \frac{VP}{VP + FN}, \quad \textit{Ecuación 6: Sensibilidad (Recall)}$$

$$\text{Especificidad} = \frac{VN}{VN + FP}, \quad \textit{Ecuación 7: Especificidad}$$

$$\text{Exactitud} = \frac{VP + VN}{VP + VN + FP + FN} \quad \textit{Ecuación 8: Exactitud (Accuracy)}$$

Donde:

- *VP*: Verdaderos Positivos (predijo "explícita" y lo era).
- *VP*: Verdaderos Positivos (predijo "explícita" y lo era).
- *FP*: Falsos Positivos (predijo "explícita" pero no lo era).
- *FN*: Falsos Negativos (dijo "no explícita", pero sí lo era).

En el apartado *B.Resultados* de esta misma sección, se interpretarán las métricas obtenidas, no sólo desde un punto de vista técnico sino también operativo. En él, se evaluarán los resultados iniciales tras el *fine-tuning,* que permitió comprobar el rendimiento base del modelo personalizado por primera vez y además una segunda evaluación tras haber aplicado un sistema de *feedback*. Ambos serán clave a la hora de estudiar la efectividad del proceso de aprendizaje supervisado y ver si la propuesta es apta para llevarla a la realidad.

5. Comparativa con GPT sin personalizar

Una parte del experimento fue realizar una comparativa entre los resultados que mostraban el modelo personalizado y los de la versión estándar de ChatGPT. Este paso se justifica por el objetivo final de que el modelo sea implementable en plataformas reales como sistema de etiquetado automático o control parental.

Para ello, se seleccionaron canciones que no se habían introducido al modelo y se pidieron a ambos modelos que clasificasen por “explícitas” o “no explícitas” y se estudió si estaban de acuerdo o no con sus respectivas predicciones. Si bien es verdad que los resultados se estudiarán en la siguiente sección, se puede adelantar que los modelos coincidieron en el 59% de los casos.

Aunque ChatGPT muestra un alto porcentaje de aciertos, es verdad que no es lo suficientemente sensible al contexto cultural y lingüístico de los géneros musicales seleccionados. Con este resultado podríamos adelantarnos a corroborar que el modelo personalizado detecta con mayor precisión el contenido sensible de las canciones puesto que es capaz de incluir incluso las referencias implícitas.

Todos los pasos descritos en esta sección tienen una correspondencia clara con los objetivos del presente Trabajo de Fin de Grado permitiendo generar evidencia empírica rigurosa que

demuestra la capacidad del modelo. Así, el diseño experimental expuesto, no sólo sigue una lógica técnica sólida, sino que también se alinea con la finalidad social del proyecto: proteger a los públicos vulnerables mediante herramientas automatizadas, éticas y adaptadas al lenguaje real de las canciones.

B. Resultados y validación de hipótesis

Como bien se exponía anteriormente, cuando se han terminado las fases de entrenamiento y *fine-tuning,* evaluar el modelo y su rendimiento es una fase esencial. Para poder llevar esto a cabo se utilizaron las métricas que se han expuesto en la metodología y que se han retirado en el diseño del experimento: precisión, sensibilidad, especificidad y exactitud y que se calcularon a partir de la matriz de confusión. Dentro del campo del *Business Analytics,* estas métricas son de suma importancia, puesto que permiten validar empíricamente si el modelo ha aprendido de forma efectiva a distinguir el contenido explícito y el no explícito. Gracias a estas métricas se han podido evaluar el porcentaje total de aciertos, la capacidad del modelo para detectar correctamente las canciones explícitas y la detección correcta de canciones no explícitas.

- Evaluación inicial y primera matriz de confusión

Para poder llevar a cabo la evaluación inicial, se compararon las predicciones realizadas por el modelo preentrenado con las clasificaciones hechas por la experta, teniendo en cuenta aspectos tanto cuantitativos como cualitativos. Así, tras el *fine-tuning,* se eligieron otras 31 canciones, que no se utilizaron durante la fase del entrenamiento, y entre ellas, la experta previamente etiquetó 16 como "explícitas" y 15 como "no explícitas". A partir de las predicciones realizadas por el modelo, se resumieron los aciertos y errores que se recogieron en la siguiente matriz de confusión:

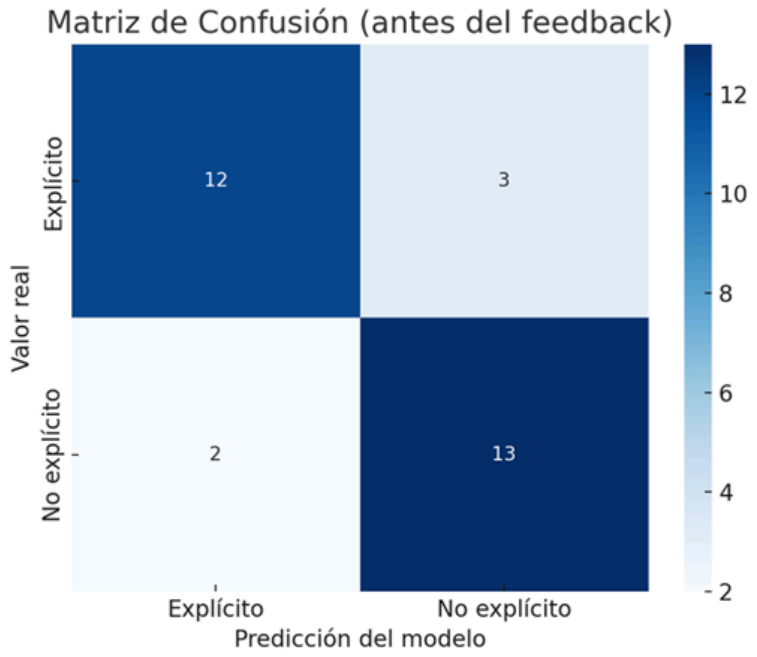

Figura 4. Matriz de confusión antes del feedback

Esta matriz demuestra que 12 de las 15 canciones explícitas fueron correctamente identificadas por el modelo, y 13 de las 15 no explícitas, resultando en 2 falsos positivos y 3 falsos negativos. De esta matriz se pudieron calcular las métricas de evaluación fundamentadas las fórmulas pertinentes y obteniendo los siguientes resultados:

$$\text{Exactitud} = \frac{VP + VN}{VP + VN + FP + FN} = \frac{12 + 13}{12 + 13 + 2 + 3} = \frac{25}{30} = \mathbf{83\%}$$

Ecuación 9: Exactitud

$$\text{Precisión} = \frac{VP}{VP + FP} = \frac{12}{12 + 2} = \frac{12}{14} = \mathbf{86\%}$$

Ecuación 10: Precisión (Precision)

$$\text{Sensibilidad} = \frac{VP}{VP + FN} = \frac{12}{12 + 3} = \frac{12}{15} = \mathbf{80\%}$$

Ecuación 11: Sensibilidad (Recall)

$$\text{Especificidad} = \frac{\text{TN}}{\text{TN} + \text{FP}} = \frac{13}{13 + 2} = \frac{13}{15} = 0{,}87 = 87\%$$

Ecuación 12: Especificidad

Estos resultados se pueden reflejar de manera visual en el siguiente gráfico de barras:

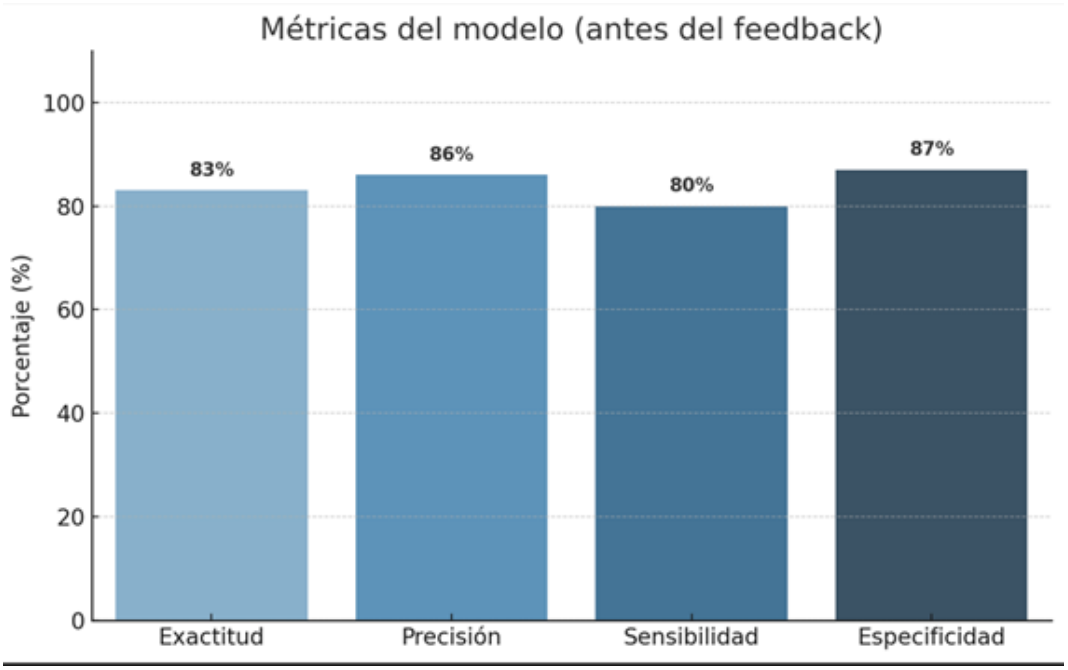

Figura 5. Métricas del modelo antes del feedback

Estos resultados permiten extraer las conclusiones acerca del funcionamiento y comportamiento del modelo: En primer lugar, en términos de exactitud, el modelo ha alcanzado un 83% de eficacia global del modelo, es decir, tiene un 83% de predicciones correctas, tanto en aquellas canciones con contenido explícito, como en aquellas con contenido "limpio". Esta cifra representa que el modelo tiene un rendimiento robusto, con una buena capacidad para generalizar sin estar sesgado hacia una sola clase, teniendo en cuenta la complejidad semántica de la tarea. Además, implica que se producen 5 errores por cada 30 decisiones, que aunque es aceptable, dentro de un entorno competitivo hay recorrido por delante.

Por otro lado, en cuanto al valor de la precisión se ha obtenido un 86%, esto indica que el modelo acierta a la hora de etiquetar algo como positivo, es decir, que predice la canción como explícita y realmente lo es. Así, el modelo solo comete errores en el 14% de los casos (falsos positivos). Los falsos positivos generan fricción, provocando la necesidad de revisar manualmente las decisiones incorrectas que toma el modelo, además esto implica que a efectos de la realidad, una canción sea etiquetada como "no apta" injustamente traduciéndose en una baja confianza en el modelo. Por tanto, aunque 86% es un valor robusto, debería priorizarse un valor más alto de precisión.

La sensibilidad del modelo en este punto es del 80%, lo que indica que el modelo acierta en ese porcentaje a la hora de detectar correctamente los casos positivos, que en este caso son las canciones que realmente tienen contenido sexual explícito. Por tanto, 8 de cada 10 canciones explícitas son detectadas por el modelo, dejando un 20% que pasa por desapercibido. Esto implica que el modelo tiene un buen filtro para detectar el contenido sensible, lo que sería útil para una futura implementación en plataformas de streaming con control parental o espacios educativos. Así, el modelo cumple razonablemente con el objetivo principal de proteger al usuario.

No obstante, no podemos pasar por alto el 20% restante, puesto que esto implicaría que hay contenido que al no ser detectado, puede llegar a usuarios sensibles por lo que hay que seguir en búsqueda de un mayor refinamiento. El tipo de lenguaje metafórico ambiguo o de jerga podría explicar este porcentaje puesto que son difíciles de captar y como destacan Chen et al. (2023), detectar expresiones sexuales implícitas o figuradas es un verdadero reto. Además, Kim & Mun (2019) también señalan que el lenguaje soez no estandarizado y que no está recogido en diccionarios convencionales, además son más difíciles de identificar.

Por otro lado, el modelo ha demostrado que tiene un valor de especificidad del 87%, lo que indica y demuestra su eficacia para descartar contenido neutro o no sexualizado. De esta manera, solo 2 de cada 15 canciones limpias fueron erróneamente censuradas. Podríamos decir por tanto, que el modelo ha aprendido a distinguir de forma eficaz expresiones neutras, coloquiales o ambiguas sin confundirlas con lenguaje explícito. Esto evita censurar o retirar del alcance de ciertos públicos canciones que no contienen expresiones sexuales explícitas. Esto implica una alta eficacia operativa y menores falsos positivos, lo que lleva a aumentar la credibilidad del sistema frente a los artistas y los usuarios, ya que pocas canciones neutras

son clasificadas de manera errónea. Por tanto, el modelo ya demuestra una madurez destacable.

Tras este profundo análisis podemos concluir que aunque la fase de *fine-tuning* ha producido resultados favorables que han sido reflejados a través de unas sólidas métricas, siguen quedando matices que mejorar para lograr un modelo que pueda implementarse en la realidad produciendo un mínimo de confianza tanto a los autores como a los usuarios. Así, aunque el modelo ha demostrado eficacia, se busca optimizar los resultados.

- **Reentrenamiento, retroalimentación y mejoras**

Tras esta primera evaluación, se decidió aplicar una estrategia de retroalimentación, también denominado *feedback loop* para mejorar el modelo. Esta decisión fue motivada por la orientación del *Business Analytics* a estar en contínua mejora, en la que evaluamos, aprendemos y ajustamos en búsqueda de una reducción de errores de clasificación. Por tanto, en base a los errores cometidos en la fase anterior, se dotó al modelo de una serie de *feedback* para que aprendiese de sus errores. Después de esto, se volvió a evaluar el modelo utilizando nuevas canciones a las que el modelo todavía no se había enfrentado. La matriz de confusión mostró los siguientes resultados:

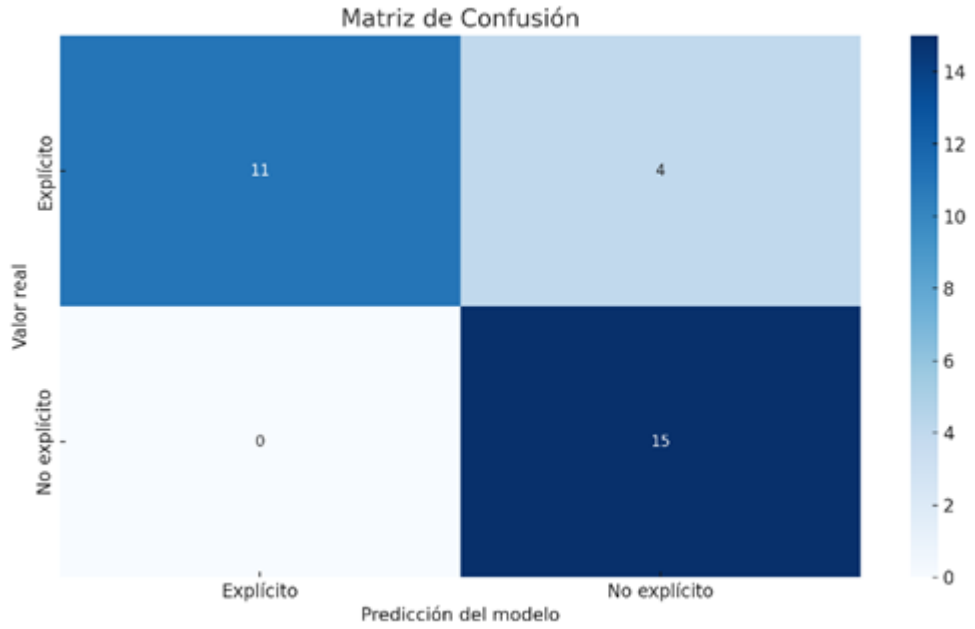

Figura 6. Matriz de confusión del modelo después del feedback

A priori, esta matriz muestra un claro avance: se han eliminado los falsos positivos por completo, lo que significa que el modelo no etiqueta ninguna canción como explícita cuando realmente no lo es. De esta matriz de confusión y aplicando las mismas fórmulas que antes del *feedback*, se pueden extraer estas nuevas métricas:

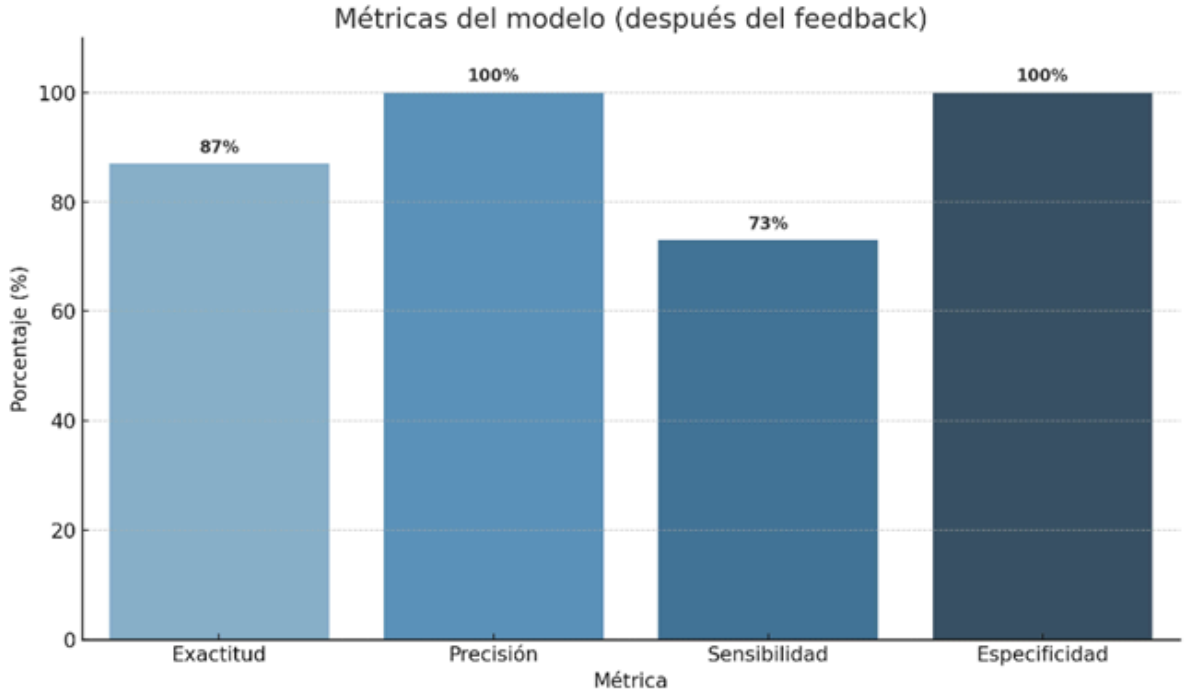

Figura 7. Métricas del modelo después del feedback

Estos resultados implican lo siguiente: en primer lugar, la precisión es del 100%, que como se comentaba anteriormente supone que el modelo no ha cometido ningún falso positivo lo cual será llamativo para aquellos artistas cuyas canciones antes se censuraban injustamente. Esto además implica que cuando el modelo afirma que una canción es explícita, verdaderamente lo es. Por otro lado, se demuestra que la sensibilidad se ha reducido al 73% . No obstante, este valor sigue siendo un valor aceptable y demuestra que tras el *feedback,* el modelo se ha vuelto más conservador priorizando evitar falsos positivos, lo cual ha conseguido al 100%.

Además, uno de los logros más importantes de haber implementado el *feedback,* ha sido lograr una especificidad del 100%, donde el modelo no marca ninguna canción no explícita como explícita, eliminando fricciones con usuarios o artistas y así aumentando la fiabilidad del modelo. Por último, la exactitud del 87% establece que el modelo acierta en 26 de 30 canciones, superando los 25 aciertos que se obtuvieron antes del *feedback.* Podemos

considerar que este resultado es consistente comparándolo con trabajos previos, como por ejemplo el de Chen et al. (2023), que logran un 96% aplicando un enfoque de aprendizaje profundo con ensemble models para detectar contenido explícito en letras de canciones. Si bien es verdad que el número de falsos negativos aumentó levemente de 3 a 4, es verdad que la eliminación total de falsos positivos impulsó gran exactitud general.

El siguiente gráfico demuestra la mejora en las métricas del modelo:

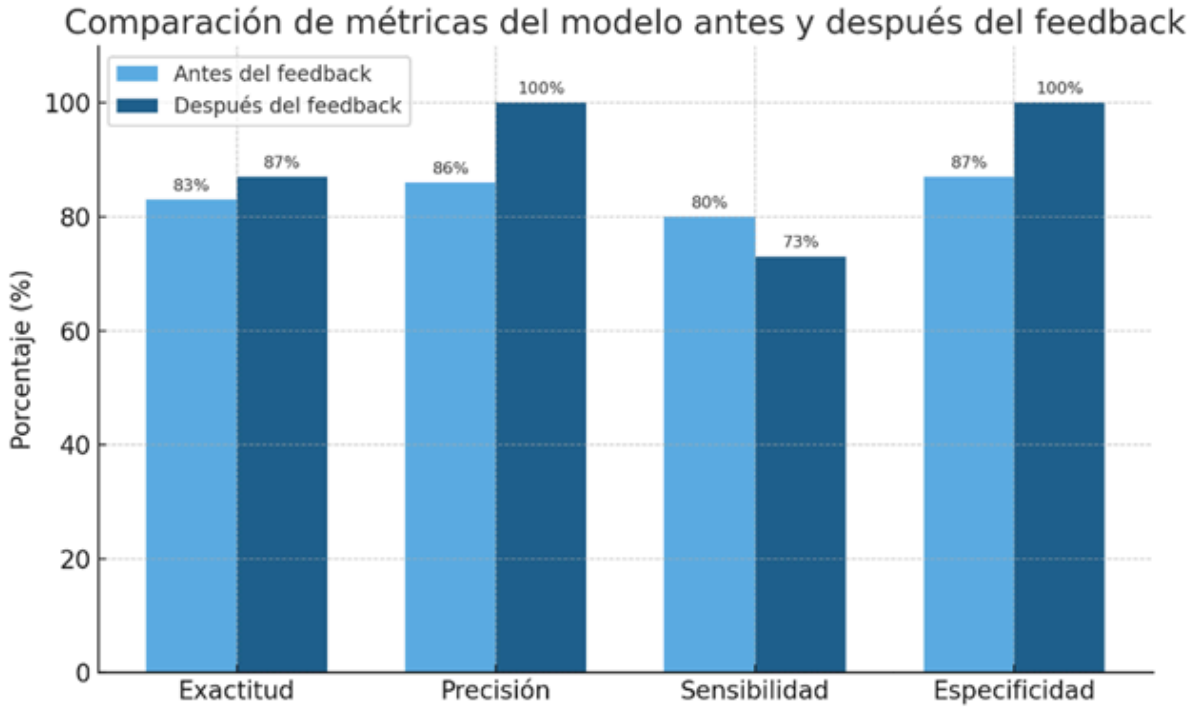

Figura 8. Comparativa de métricas del modelo antes y después del feedback

Así, el *feedback* ha resultado ser una estrategia eficaz pues el modelo ha mejorado su equilibrio entre rigidez y permisividad reduciendo así los errores operativos y mejorando la confiabilidad global del sistema. Si bien es verdad que se ha sacrificado la sensibilidad, ha sido en virtud de obtener mejoras en otras métricas. El modelo ajustado es más robusto, fiable y apto para su implementación en plataformas reales como Spotify o Apple Music

- **Comparativa con modelo GPT sin personalizar**

El hecho de comparar el rendimiento entre el modelo personalizado y uno sin personalizar-ChatGPT- fue una de las claves para analizar y validar las hipótesis que se establecieron al principio del trabajo.

Para llevar a cabo la comparativa entre ambos modelos se cogieron 50 canciones que no se utilizaron a la hora de entrenar el modelo y se solicitó a los dos modelos que lo clasificasen en "explícito" o "no explícito". Desde ahí y con los resultados de ambas predicciones se realizó el cálculo del porcentaje en el que estaban de acuerdo (*ground truth*) y que se pueden ver en el siguiente gráfico:

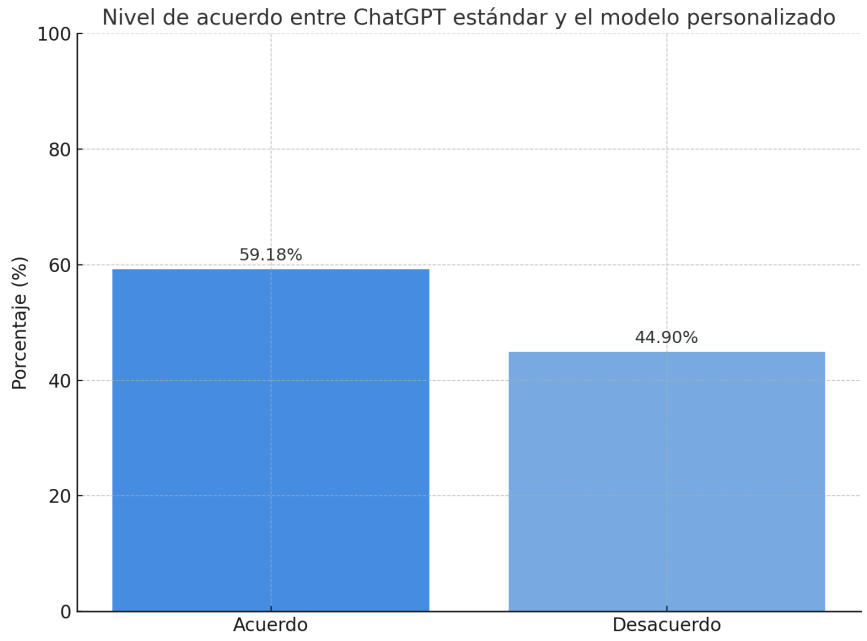

Figura 9. Nivel de acuerdo entre ChatGPT estándar y el modelo personalizado

Tal y como se demuestra, los resultados establecieron que el modelo personalizado concordaba en el 59,1% de los casos con la clasificación humana experta pero la versión sin personalizar estaba en desacuerdo en el 44,9% de los casos. Estas cifras demuestran como la diferencia es significativa puesto que el modelo fine-tuned es más sensible a la hora de detectar el contenido explícito y entiende mejor el lenguaje, expresiones y dobles sentidos del reguetón y el trap.

Estos resultados y conclusiones demuestran cómo los modelos ajustados serían más precisos que los modelos tradicionales, pero sobre todo serían más aptos para generalizar sobre las características propias de los géneros musicales que se están estudiando. ChatGPT si bien está cada vez más actualizado y entrenado, todavía no entiende el contexto de fondo y la jerga que se utiliza en este ámbito y hace que este modelo no personalizado aplica criterios demasiado amplios o neutros. Mientras, el modelo personalizado logró captar los matices culturales y lingüísticos que son esenciales para una clasificación fiable.

Todos los resultados obtenidos y estudiados a lo largo del experimento confirman que el modelo personalizado para este trabajo es sólido y a su vez valida las hipótesis planteadas. Además este rendimiento también es comparable con el de otros enfoques recientes, como por ejemplo, el de Rospocher (2020), quien alcanzó un F1-score del 88% utilizando embeddings enriquecidos a nivel de subpalabra para la detección automática de letras explícitas. Por todo esto, además podemos confirmar que el aprendizaje supervisado es eficaz a la hora de adaptar modelos generalistas a tareas más específicas como aquellas más sensibles o ambiguas.

En cuanto a la hipótesis nula (H_0), que planteaba que los métodos tradicionales que usan diccionarios y modelos estadísticos simples serían más eficaces, tras este profundo análisis y el estudio de los resultados, podemos confirmar que queda descartada. La capacidad que tiene el modelo personalizado de comprender el contexto y la semántica, claramente supera a cualquier otro modelo basado en reglas simples y estáticas.

Por otro lado, en cuanto a la hipótesis alternativa (H_1) que establecía que un modelo basado en transformers personalizado mediante fine-tuning alcanzaría un mayor rendimiento, queda demostrada empíricamente no sólo por los resultados que se han obtenido en la evaluación inicial sino que queda superada con los resultados obtenidos tras el *feedback.*

Por todo esto podemos concluir que este trabajo no solo ha conseguido entrenar un modelo de lenguaje personalizado para detectar contenido explícito, sino que se ha abordado la importancia de dar respuesta a la problemática social.

6. CONCLUSIONES Y TRABAJO FUTURO

A lo largo del presente Trabajo Fin de Grado se ha abordado el problema social y su solución tecnológica: la importancia de que la inteligencia artificial responda a regular la exposición de los más jóvenes a las canciones con contenido sexual y explícito, especialmente en los géneros de trap y reguetón. Estos estilos musicales, principalmente consumidos por niños y adolescentes, suelen transmitir mensajes que cosifican el cuerpo y banalizan relaciones afectivas o sexuales, incidiendo en los valores, actitudes y conductas que asumen los jóvenes.

Este es un problema de carácter sociocultural que nos afecta a todos: la generación del "mañana" está en peligro, y frente a esto, en este trabajo se ha planteado una solución innovadora que combina los avances tan recientes de la inteligencia artificial- mediante modelos de lenguaje basados en transformers-. Así, el modelo planteado, ya ha sido previamente entrenado de manera general y el trabajo ha supuesto su refinamiento a través de técnicas de *fine-tuning* para que cumpla una tarea muy específica: la detección y clasificación automática de canciones con contenido explícito y sexual. Para ello se han seguido unos pasos muy pautados y cuidados consistentes en el desarrollo y curación de una base de datos propia compuesta por canciones previamente etiquetadas por una experta, y posteriormente una evaluación rigurosa del comportamiento del modelo.

Esta solución tecnológica ha permitido cumplir con el objetivo general que se propuso en el apartado de "Alcance del trabajo": desarrollar un sistema automatizado avanzado de lenguaje natural para detectar el contenido explícito en canciones de reguetón y trap. Además, se podría decir que también se ha cumplido con el objetivo 1 puesto que queda más que demostrado que el modelo personalizado de GPT puede ser útil a la hora de detectar automáticamente contenido sensible.

Además, a través de un elaborado corpus y la elaboración de tablas que establecen las palabras y frases explícitas de cada canción del corpus, también se ha conseguido el segundo objetivo: que el modelo esté adaptado al lenguaje del reguetón y trap. Asimismo, el objetivo 3 referido a que el modelo cumpla con un mínimo rendimiento, ha quedado también demostrado con las métricas de precisión, sensibilidad, especificidad y exactitud. Por último, el objetivo 4 queda parcialmente cubierto tras en la propuesta de líneas de trabajo futuro se establece la aspiración de que el sistema sea integrado en plataformas musicales reales

Lo realmente valioso en este trabajo no es solo la elaboración de un modelo capaz de llevar a cabo esta tarea, sino el enfoque del Business Analytics que consiste en conectar la tecnología con un problema real. De hecho, todo el trabajo se ha analizado desde la perspectiva de la analítica de datos: desde su calidad, hasta la evaluación de resultados y la justificación de las decisiones metodológicas adoptadas. No solo se trata de que el modelo funcione sino de entender para qué sirve y qué ventajas puede traer su implementación en la vida real.

En cuanto a la validación de hipótesis, como bien ha quedado establecido en la sección anterior, tras el profundo análisis y evaluación empírica, queda rechazada la hipótesis nula (H_0) puesto que se ha demostrado que los modelos basados en transformers son más eficaces que los tradicionales. Además, la hipótesis alternativa (H_1), también ha quedado empíricamente demostrada tras analizar las métricas antes y después del *fine-tuning* y del *feedback.*

Además, el análisis desde este enfoque ha permitido centrar la clave en conectar los datos con decisiones estratégicas, sociales y éticas, alineándose a la perfección con uno de los pilares fundamentales del Business Analytics. Esta perspectiva nos permite cambiar nuestra mirada: la inteligencia artificial no solo es una herramienta técnica sino un recurso que puede ser utilizado en favor de la sociedad, para transformar sectores y evolucionar con ellos. Así, este trabajo proporciona la intersección entre los datos, la inteligencia artificial y la responsabilidad social.

Con este trabajo a la espalda, podemos mirar al futuro y encontrarnos con múltiples líneas de desarrollo e investigación. Para comenzar, sería muy interesante ampliar el corpus en cuanto a diversidad de subgéneros, países, lenguajes y expresiones culturales. De esta manera el modelo podrá traspasar fronteras y extenderse a nuevas tendencias musicales.

Por ello, este modelo no debe entenderse como un fin en sí mismo sino como un medio para regular la manera en la que interactuamos con la música que tanto al día está.

Desde la perspectiva del *Business Analytics* se ha demostrado como la transformación de datos crudos pueden llegar a ser piezas de conocimiento de gran valor y que pueden llegar a convertirse en aplicaciones tangibles. Así, la combinación de datos textuales con un modelo de lenguaje natural puede responder a problemáticas reales. Así, este trabajo ha conseguido uno de los elementos claves de *analytics:* conectar el análisis con problemas concretos a través de un tratamiento adecuado de los datos y una comprensión profunda del texto.

Desde la perspectiva del negocio, no hay duda alguna de que este modelo es una oportunidad para empresas tecnológicas y plataformas como Spotify o Apple Music y que tienen que estar constantemente evolucionando y adaptándose para poder ser líderes en esta era digital. Su incorporación de modelos largos de lenguaje mejora además la percepción de la marca puesto que fortalece su compromiso ético.

Así, podrían también utilizar estas tecnologías para por ejemplo la personalización de playlists que recomienda automáticamente canciones seguras para menores que han sido filtradas previamente. Además se podrían generar una serie de alertas antes de reproducir canciones con contenido explícito para poder integrarlo en entornos familiares a través de controles parentales inteligentes. Su implementación también puede favorecer a las estrategias de segmentación de mercado basadas en valores, creando así perfiles de usuarios sensibles o perfiles "Kids". Esto, en términos de ROI, supone además un valor diferencial frente a la competencia dentro del sector de plataformas de música, puesto que ninguna ha incorporado todavía capacidades de moderación automatizada que reduzcan los riesgos de difusión de contenido inadecuado o sensible.

En definitiva, este Trabajo de Fin de Grado demuestra como el *Business Analytics* puede responder a los problemas de la realidad y generar soluciones con gran impacto siendo un motor de cambio en nuestra sociedad.

Pese a todos estos resultados y descubrimientos, también se han encontrado limitaciones relativas tanto al enfoque como al contexto del problema, coincidiendo en parte con las planteadas en las restricciones en la parte inicial del trabajo. Se ha confirmado la dificultad de construir un corpus equilibrado, siendo complicada la búsqueda de canciones de reguetón y trap que no fueran explícitas. También, en la construcción de este corpus, se encontró un grado de dificultad de establecer unas medidas objetivas de qué contenido se puede considerar explícito o no, puesto que es una tarea llena de subjetividad.

Este trabajo no se queda aquí, sino que es simplemente el principio. Se ha logrado asentar una serie de bases sólidas para crear un sistema automatizado que identifique cuáles canciones son explícitas y cuáles no. Esto deja la puerta abierta a la expansión y la mejora. Para empezar se podría ampliar la base de datos (sin sobre ajustar el modelo) para que tenga una mayor comprensión del lenguaje de estos géneros musicales. También, desde un punto de vista algorítmico, se podría optimizar la especificidad del modelo cuidando la sensibilidad

para poder evitar falsos positivos. Además se podría explorar el modelo con otros modelos más ligeros (computacionalmente) como Llama o DistilGPT para poder integrar el sistema en dispositivos móviles.

El mayor de los objetivos a futuro sería integrar este modelo dentro de un pipeline real de despliegue en plataformas de contenido musical como Spotify o Apple Music a través de APIs para poder analizar canciones en tiempo real y que además se pueda expandir a nuevos sectores y lenguas de la música.

En conclusión, este Trabajo de Fin Grado ha materializado una solución tecnológica basada en modelos de lenguaje que responde a una necesidad real: proteger a los público más vulnerable de aquellas canciones que pueden comprometer sus valores. Si bien queda un buen trabajo por delante, este trabajo presenta un primer paso hacia una música más consciente con consumidores más informados.

REFERENCIAS

Addanki, S., & Murthy, N. (2022). *Text content moderation model to detect sexually explicit content*. *CS230: Deep Learning, Stanford University*.

Bhatti, A. Q., et al. (2018). *Explicit content detection system: An approach towards a safe and ethical environment*. *Applied Computational Intelligence and Soft Computing*. Fell, M., et al. (2019). *Comparing Automated Methods to Detect Explicit Content in Song Lyrics*. *RANLP*.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., … & Amodei, D. (2020). *Language models are few-shot learners*. Advances in Neural Information Processing Systems, 33, 1877-1901. https://doi.org/10.48550/arXiv.2005.14165

Chen, X., Aljrees, T., Umer, M., Karamti, H., Tahir, S., Abuzinadah, N., … & Ashraf, I. (2023). *A novel approach for explicit song lyrics detection using machine and deep ensemble learning models*. PeerJ Computer Science, 9, e1469. https://doi.org/10.7717/peerj-cs.1469

Chin, H., Kim, J., Kim, Y., Shin, J., & Yi, M. Y. (2018). Explicit content detection in music lyrics using machine learning. *2018 IEEE Intl. Conf. on Big Data and Smart Computing (BigComp)*, 517-521. https://doi.org/10.1109/BigComp.2018.00101

Clérice, T. (2023). *Detecting sexual content at the sentence level in first millennium Latin texts*. *arXiv*. Colmenares-Guillén, L. E., & Jiménez-Aguilera, J. L. (2023). *Una aproximación para la detección de contenido sexual en conversaciones digitales*. *CienciAmérica*.

Darroch, K., & Weir, G. R. S. (2014). *Measuring Sexually Explicit Content in Text Documents*. *Cyberforensics 2014*.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of deep bidirectional transformers for language understanding*. Proceedings of NAACL-HLT 2019, 4171-4186. https://doi.org/10.48550/arXiv.1810.04805.

Diario Libre. (2021, 19 de noviembre). *Expertos en conducta humana alertan a diputados sobre impacto de contenido explícito en música urbana*. https://www.diariolibre.com/actualidad/politica/expertos-en-conducta-humana-alertan-a-diputados-sobre-impacto-de-contenido-explicito-en-musica-urbana-AH30041414

Díez-Gutiérrez, E. J., & Muñiz-Cortijo, L. M. (2023). Educación reguetón: ¿Educa el reguetón en la desigualdad? *Perfiles Educativos, 45*(179). https://doi.org/10.22201/iisue.24486167e.2023.179.60295

Kim, J., & Mun, Y. (2019). *Offensive content detection in lyrics using deep neural networks*. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, 137–142. https://doi.org/10.18653/v1/P19-2024

Markov, T., et al. (2023). *A Holistic Approach to Undesired Content Detection in the Real World*. *AAAI*.

Martino, S. C., Collins, R. L., Elliott, M. N., Strachman, A., Kanouse, D. E., & Berry, S. H. (2006). Exposure to degrading versus nondegrading music lyrics and sexual behavior among youth. *Pediatrics, 118*(2), e430-e441. https://doi.org/10.1542/peds.2006-0131

Milenio. (2020.). *¿Puede el reguetón inducir a la violencia?* https://www.milenio.com/espectaculos/musica/puede-el-regueton-inducir-a-la-violencia

Molpeceres Barrientos, G., et al. (2020). *Machine Learning Techniques for the Detection of Inappropriate Erotic Content in Text*. *International Journal of Computational Intelligence Systems*.

Povedano Álvarez, D., et al. (2023). *Learning Strategies for Sensitive Content Detection*. *Electronics*Gutfeter, W., et al. (2024). *Detecting sexually explicit content in the context of child sexual abuse materials (CSAM)*. *arXiv*.

RIAA. (2021). *Parental Advisory Label*. Recording Industry Association of America. https://www.riaa.com/resources-learning/parental-advisory-label/

Rospocher, M. (2020). Explicit song lyrics detection with subword-enriched word embeddings. *Expert Systems with Applications, 163*, 113749. https://doi.org/10.1016/j.eswa.2020.113749.

Sanz Torres, Í. (2023). *Detección de Contenido Sexual mediante Aprendizaje Profundo y Aprendizaje por Transferencia*. *TFG, UCM*.

Spotify. (2022, 30 de noviembre). *Ya están aquí: los artistas, canciones y podcasts más escuchados, además de las tendencias de escucha del 2022*. Spotify Newsroom. https://newsroom.spotify.com/2022-11-30/ya-estan-aqui-los-artistas-canciones-y-podcasts-mas-escuchados-ademas-de-las-tendencias-de-escucha-del-2022/

Okulska, I., & Wiśnios, E. (2023). *Towards Harmful Erotic Content Detection through Coreference-Driven Contextual Analysis*. *arXiv*

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … & Polosukhin, I. (2017). *Attention is all you need*. Advances in Neural Information Processing Systems, 30, 5998-6008. https://doi.org/10.48550/arXiv.1706.03762.

Vázquez Benito, A. O. (2023). *Desarrollo de un sistema de software basado en Transformers para la detección de lenguaje de odio en medios sociales en español*. *Instituto Tecnológico Superior de Teziutlán*

Yu, Y., & Yin, X. (2023). *A hypersensitive intelligent filter for detecting explicit content in learning environments*. *Journal of Web Engineering*.

ANEXO

1. ECUACIONES

$$H_0 : \mu_t - \mu_c \leq 0$$

Ecuación 1: Hipótesis nula (H_0)

$$H_1 : \mu_t - \mu_c > 0$$

Ecuación 2: Hipótesis alternativa (H_1)

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Ecuación 3: Mecanismo de atención escalar

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_H)W^O, \quad \text{donde} \quad \text{head}_i = \text{Atención}(QW_i^Q, KW_i^K, VW_i^V)$$

Ecuación 4: Atención multi-cabeza (Multi-Head Attention)

$$\text{Precisión} = \frac{VP}{VP + FP}$$

Ecuación 5: Precisión (Precision)

$$\text{Sensibilidad} = \frac{VP}{VP + FN}$$

Ecuación 6: Sensibilidad (Recall)

$$\text{Especificidad} = \frac{VN}{VN + FP}$$

Ecuación 7: Especificidad

$$\text{Exactitud} = \frac{VP + VN}{VP + VN + FP + FN}$$

Ecuación 8: Exactitud (Accuracy)

$$\text{Exactitud} = \frac{VP + VN}{VP + VN + FP + FN} = \frac{12 + 13}{12 + 13 + 2 + 3} = \frac{25}{30} = \mathbf{83\%}$$

Ecuación 9: Exactitud

$$\text{Precisión} = \frac{VP}{VP + FP} = \frac{12}{12 + 2} = \frac{12}{14} = \mathbf{86\%}$$

Ecuación 10: Precisión (Precision)

$$\text{Sensibilidad} = \frac{VP}{VP + FN} = \frac{12}{12 + 3} = \frac{12}{15} = \mathbf{80\%}$$

Ecuación 11: Sensibilidad (Recall)

$$\text{Especificidad} = \frac{\text{TN}}{\text{TN} + \text{FP}} = \frac{13}{13 + 2} = \frac{13}{15} = 0{,}87 = 87\%$$

Ecuación 12: Especificidad

2. FIGURAS

Figura 1. Problemática de la detección del contenido explícito

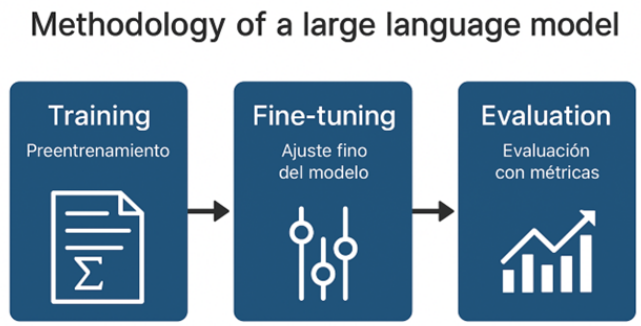

Figura 2. Fases del desarrollo de un modelo grande de lenguaje (LLM)

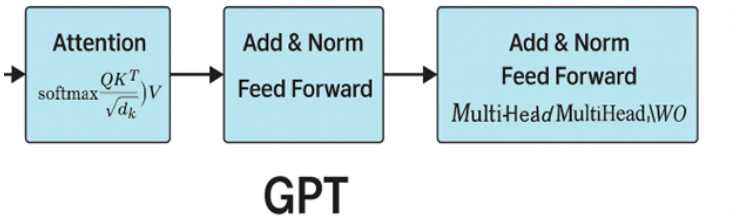

Figura 3. Estructura interna de una capa del modelo GPT

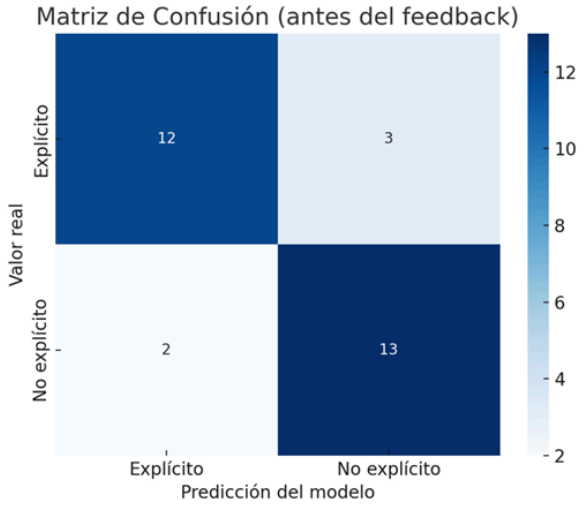

Figura 4. Matriz de confusión antes del feedback

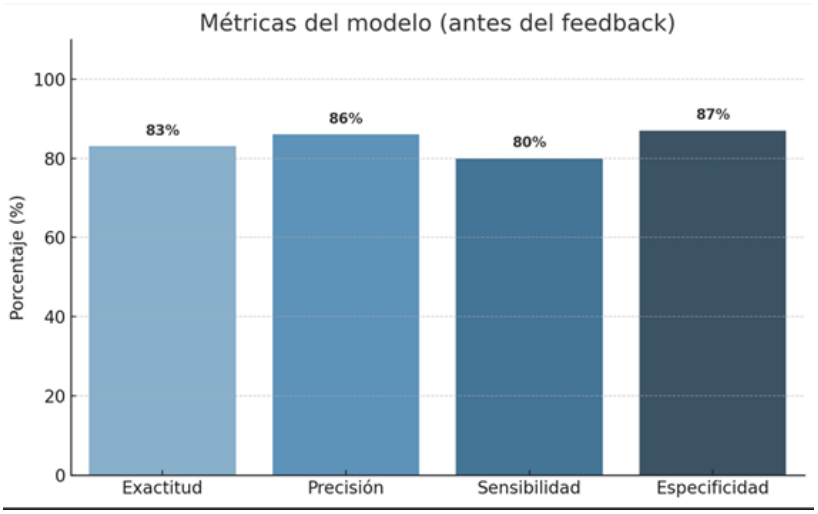

Figura 5. Métricas del modelo antes del feedback

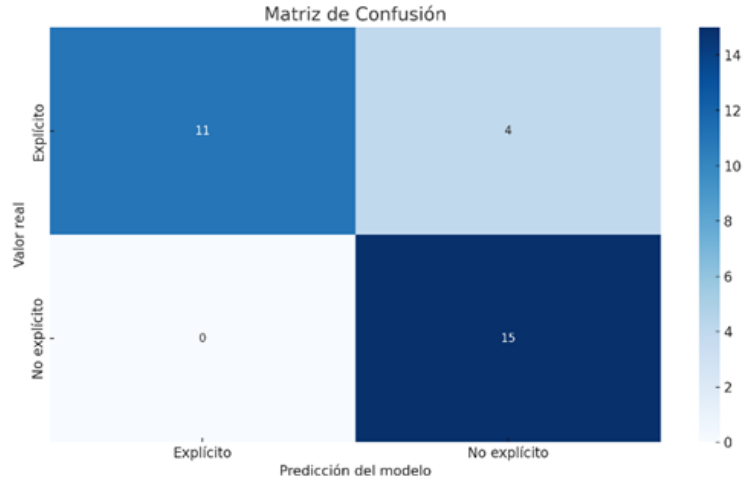

Figura 6. Matriz de confusión del modelo después del feedback

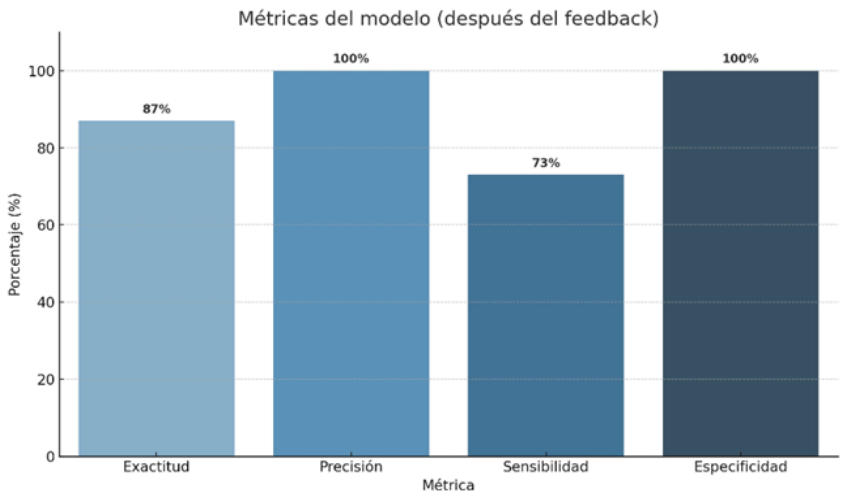

Figura 7. Métricas del modelo después del feedback

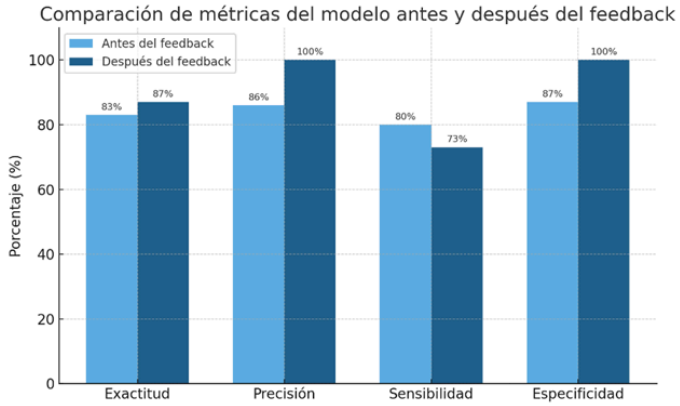

Figura 8. Comparativa de métricas del modelo antes y después del feedback

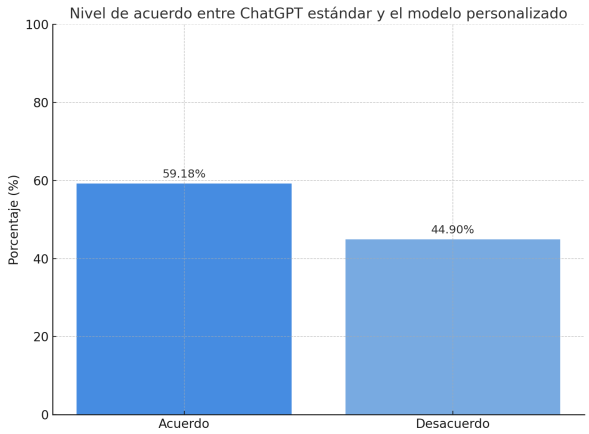

Figura 9. Nivel de acuerdo entre ChatGPT estándar y el modelo personalizado

3. TABLAS

Tabla 1

Canción	Frases Explícitas
China	Borracho y prendío' ; 'Yo estaba contigo perreando' ; 'Y se está prendida en fuego' ; 'Ese culo es una amenaza' ; 'Ese booty que hasta un ciego puede ver'
Bandido	Condone´de colore; la parto a lo crayola'; Si pone' el booty en reversa
Ignorantes	Pero que rico cuando chingamo' ; 'Te quitas el panti' ; 'Y pasas las notas en mi cama'
La Modelo	Pa' que te mojes como yo me mojé' ; ' Quiero hacerte tantas cosas' ; 'Quiero probar algo de ti' ; 'Me mete sudando, su cuerpo de mora'
Pa' Romperla	Hoy quiero que la noche salga, pa romperla' ; 'Fumando marihuana' ; ' Te voy a meter el peine full, como pistola de bichote' ; 'Quiero que juegues conmigo' ; 'Baby, póngase de espalda Pa' romperla' ; ' Conmigo es que tu baby se pone bellaca'
Fantasías	Voy a tocarte to'a, ¿qué tal si se nos da?' ; 'Susurrando al oído te empiezas a calentar' ; 'Terminamo en mi cama' ; 'yo clantandome tu calentandote'
Secreto	Te beso el cuello y las piernas también' ; 'Y yo solo en mi cama, Dios es el testigo' ; 'siempre nos comemos' ; 'Y te hago el amor bien rico' ; 'Pero, pa'l sexo, es temprano' ; 'Y siempre lo hacemos rico en mi cuarto, bebé'- referencia a actividad sexual
Hasta Que Dios Diga	Te lo hago hasta que Dios diga' ; ' Y ese traje se ve bien con tus nalgas apretas' ; 'Tú estas prendida'; 'Tú desnuda en mi cama' ; 'Anoche te soñé Y me quedé con las gana' ; 'te quiero disfrutar' ; 'Te vo'a dar hasta que Dios diga' ; 'Tú llega' y estas putas empiezan a envidiarte' ; 'Y eso' labio' allá abajo 'tán jugoso' ; 'me tropecé con tu culito' ; 'Morderte y lamberte to'as tus parte' ; 'Hablándote al oído grosería' ; 'en el sexo tú eres mía' ; 'Ese culo está cabrón' ; 'Te toqué y estaba' humedecida' ; 'Pa' chingar, no hay que estar junto' de por vida' ; 'Ella se lo traga y me lo escupe' ; 'Tú quiere' comerme' ;
Cuatro Babys	Chingan cuando yo le digo' ; 'Polvo corrido, siempre me dan tres' ; 'en la cama, maltratarme' ; 'Me tienen bien, de sexo me tienen bien' ; 'Se encojona si se lo echo afuera' ; 'Y me paga pa' que se lo hunda' ; 'De chingar, ninguna se enzorra' ; 'Me tiene enamorado ese culote' ; 'Tengo una chiquitita nalgona' ; 'Me dice: Papi, vente adentro, sí, me preña' ; 'A todas, yo quiero darle' ; 'Y es que todas maman bien' ; 'Pero cuando chingan Gritan todas por iguales
Mujeres	Esta noche está pa bailar, beber, joder hasta que no pueda más' ; 'Desperté con una loca me dijo que era mi mujer' ; 'Y cuando sé ennota, conmigo lo goza'
Diles	Vamonos al cuarto polvo' ; 'Si les preguntan por que pal sexo yo soy tu fav' ; 'Que yo me se tus poses favoritas' ; 'Pa' hacerte venir' ; 'Que te hablo malo y que eso te excita' ; 'Yo a ti te martillo' ; 'Porque te gusta como te doy' ; 'Quiero hacerte cositas que nunca te han hecho' ; 'ven y mama' ; 'Voy a darte duro hasta por la mañana' ; 'Le abro las piernas en el balcón'; 'Hago susurrar todas tus voces por mi cuello' 'Poses abajo, arriba, Poses en cuatro, encima o de la'o, Caliente, caliente' ; 'Me dio un blowjob mientras conducía'
Otro Trago (Remix)	Otro trago y nos vamos pa'l cuarto' ; 'La dejaron sola y rompe la batidora, un perreo sin censura'

No Me Conoce (Remix)	Pero de noche conmigo le gusta portarse mal' ; 'Pero en mi camase volvio un vicio' ; 'En el fondo ella quiere más, no le gusta conformarse' ; 'Está puesta pa' bellaquear' ; 'Pero en mi cama se volvió un vicio como la 5-12' ; 'Tiene el booty XL, pero usa los panties small' ; 'La baby está muy dura' ; 'Pero en mi cama, se lo metí en 4 Y en toditas las poses' ; 'Me la como entera' ; 'Me llama pa' que yo la pruebe y cuando yo la toco, eso llueve ahí. Ella se vistió y yo la desvestí' ; 'Como cuando yo le di en todas las poses' ; 'Me tira pa' que yo la pruebe Se pone olorosa y me gusta cómo huele' ; 'hoy quiere joder' ; '
La Jeepeta (Remix)	Fumando marihuana en la jeepeta' ; 'Que tiene grande' las teta' (las teta') Quiere que yo se lo meta' ; 'Por qué no hacemos una porno como Ozuna? Bebé, dame ese culo, por fa', quítame la hambruna Es que yo como toto, por eso es que Anuel no ayuna' ; 'pa' que te mojes' ; 'Pero la que chicha siempre trae los condones' ; 'Esas tetas son un monumento (monumento) Y esto es un perreo' ; 'fumando hachís Estoy tan arrebatao' ; 'la imagino en cuatro Más de 24, en el sexo un bachillerato, y yo Si dice que no fumo, es un teatro Se toca y me manda los retratos' ; 'dentro de ella te lo hago' ; 'Dice que me espera desnuda en el hotel San Juan' ; 'Fumar hasta que tu mente se borre' ; '
Bubalu	Estoy puesto pa' ti, ¿qué tal si se nos da?' ; 'La cama hace tu-tu-tu' ; 'Ese culito en el jetski en Montego Bay' ; 'Sexing, cuddling' ; 'make you come kulosa' ; '
Quiero Repetir	Como tu lo hace otra no hya' ; 'Te lo quiero hacer otra vez' - contenido sexual explícito
Loco Contigo	Eres el número uno, y te quiero en mi cama' ; 'Con un booty fuera de lugar' ; 'Enseñando todo mucho por fuera'
Mala Mía	Lo que pasó entre tú y yo lo dejamos en la cama'
Ella Quiere Beber (Remix)	Cuando una mujer decide ser mala y no quiere dueño' ; 'Ella ahora va a vengarse en mi cama' ; 'Tu cuerpo y tu voz a mi me excitan' ;'Se emborracha y no quiere enamorarse' - referencia al alcohol
Te Boté (Remix)	Ya te olvidé y te boté' ; ' Como yo te lo hacia aquella vez' ; 'Lo que pide es un perreo sucio en la placita'
Bellaquita (Remix)	Te gusta que te den bellaqueo' ; ' A mi me gusta cuando baja downtown' ; 'Como ese culo me tiene enviciao'
Noche De Entierro	Hoy se bebe y se chinga' ; 'Vete que yo soy un perro veintucatro siete' ; 'Acariciame el petete' ; 'Y voy deborando todas las gatas que yo veo sola'
Adicto	Siempre estoy pa' ti cuando tú quieras sexo' ; 'Soy adicto a tu parte' ; 'Si es temprano, nos pegamo' a la pared' ;
X	Aqui solo subimo mami, tu que haces bajando' ; 'por la noche yo le vo'a dar'
Qué Pretendes	Yo sé que tú quieres más' ; 'Por ti me meti la pastilla' ; 'Tú lo que quiere eh joder' ; 'Me va' a ver con una y te va a joder'
Gata Only	Yo quiero chingar' ; 'Mueve los chachetes al ritmo del TikTok' ; 'Otra noche dándote, dándote. Tocándote, calentándote. Encima viniéndote. Al oído, gimiéndome' ; 'Allá abajo mojarte toa' las parte' ;
Santa	Fumando todo el día' ; 'Rehunsando nuestro cuerpo' ; 'A ti yo te rezo, mi santa' ;

Tusa	Se cansó de ser buena, ahora es ella quien los usa' ; 'Está dura y abusa'
Safaera	Si tu novio no te mama el culo, pa' eso que no mame' ; 'Me chupa la lollipop' ; '¿Cómo te atreve', mami, a venir sin panty?' ; 'Hoy se fuma como un rasta' ; 'Chocha con bicho, bicho con nalga' ; 'Te-Te está rozando mi tetilla' ; 'Las nalga' bien grande' ; 'Yo quiero perrearte'
Yonaguni	Quiero tenerte encima de mí' ; 'Y yo loco por tocarte' ; 'Shorty, tiene' un culo bien grande'
Baila Baila Baila (Remix)	Tú me tienes loco con esas caderas' ; 'llégale al código con el perreo sólido' ; 'Acércate al perímetro y rompe lo'
Zorra	Soy una zorra y me gusta el bellaqueo' ; 'Tú la jodiste con todas nosotras' ; ' jodiste con to'a' ; 'Tu hijo es una zorra' ; 'Decías que extrañaba' cómo te tocaba'
Yo Perreo Sola	A ella le gusta que le den duro y se la coman' ; 'Ella perrea sola'
Dákiti	Tú me tienes juqueao' ; 'Tú está' bien suelta' ; 'A mí sin cojones' ; 'Tú mueve' el culo fenomenal' ; 'Pa' yo devorarte como animal' ; 'Tú está' bien suelta'
Otra Noche	Otra noche más pidiéndote más' ; 'Morir chingando después de prender' ; 'Moviendo los cuadrito' ; 'Si quiere, hoy mismo se lo coloco' ; 'El corazón en dos, yo a ella en cuatro la rompo' ; 'Y si me tiene adentro, la hago sentir viva'
Se Me Olvida	Cuando estamos solos, siempre pasa lo mismo' ; 'Sigue entrenando en el body y nunca te dije: Sorry' ; 'Te hubiera echa'o tres más' ; 'Y que si no fuera por mi culpa, estaríamo' chingando ahora mismo'
1000Cosas	Me encanta cuando estás encima de mí' ; 'Comerte, con la boca, desnudarte' ; 'Sabes que esta noche los dos haremos mil cosa' ; 'Tú quieres comer-mer-mer-mer-merme Y yo chinga-ga-ga-garte, Quiero ve-ve-ve-ve-verte y desnudarte, Porque hoy quiero volver a hacerte mil cosa', Y recordar lo rico que fue' ; 'El cristal empaña'o, más humedad que el Amazona' ; 'Nadie lo hace como lo haciamo' Terminábamo' y repetíamo' El mantel ya está en la mesa pa' que nos comamo' como lo hacíamo'
Cuando Te Vi	Me obsesiona comerte otra vez' ; 'Me lo hiciste tan rico que me cuesta no volverte a llamar' ; 'Ando en busca de una turra y si es de zona sur, mejor' ;
Eskeleto	Nos fuimos pa' la cama sin ropa' ; 'Tenemos una playlist conjunta solo pa' meter las canciones de cuando chingamo' ; 'Si quieres lo meto pero no te prometo tan dentro que siento hasta el esqueleto, adentro de tu selva, húmedo, nunca está seco' ; 'Si está en sus día' del mes, el bigote me lo deja como el pelo de Ed Sheeran' ; 'Penetro por el baseline' ; 'Adentro de tu cueva, húmedo, nunca está seco'
Adivino	Y te lo quiero hacer, baby, hasta que amanezca' ; 'I just wanna fuck you,' ; 'We should have sex, sex, sex and we be daily, daily'
Qlona	Pa' que se te mojen las piernas' ; 'Qué hijueputas ganas tengo de besarte' ; 'te imaginé sin ropa' ; 'cómo te ves de culona' ; 'Nos vamos de guayeteo, fumeteo en la disco, mero perreo' ; 'Te pusiste minifalda pa' ver si yo te gateo' ; 'No te voy a mentir, no paro 'e imaginarme tu culo en tanga' ; 'Te pongas caliente y todo el cuerpo arda, arda' ; 'Y ese culito blanco, el Sol te lo ponga moreno' ; 'Y ese culito blanco, el Sol te lo ponga moreno, Lo tiene grandote' ; 'Estás provocándome aunque lo haces sin querer'

Mi Luz	Quiero hacerte cosas malas' ; 'Yo estoy con hambre y tú estás pa' comerte, yeah' ;
Friki	Quiero tocar tu cuerpo completo' ; 'Tú que andas calientica' ; 'Eso allá abajo está sticky' ; ' ni un polvo se le escapa' ; 'Me la pone' fuerte' ; 'No me excite', desnúdate, Que hoy yo te la vo'a meter'
Goteras	Fumando marihuana, no hay problema' ; 'Duermo haciéndote el amor' ; 'Encima de mí hace tiempo que no 'tás, me acordé de esa postura' ; 'Como Shakira, el culo grande del short se le sale' ; 'Pensando en cómo solita te toca' ; 'Solo yo con verla eso se pone roca' ; 'Me-meto la llave en ese toto pa' que arranque' ; 'voy pa' allá abajo tú gime' exótico'
Badgyal	Yo no soy tuya, ni de nadie' ; 'Que está buenísima' ; 'Me pone ojito’ de gata cuando baila' ; 'Y le quiere dar hasta el otro día' ; 'Perreando' ; 'Ven, a ese culo le rezo' ; 'Y con la amiga que está caliente' ; 'Mueve el culo' ; 'tremenda putonga' ; 'quiere bicho' ; 'Se me pone perra' 'Está en racha y borracha, En falda, asomando las cacha'- contenido relacionado con empoderamiento sexual
La Falda	Fuma y no le importa nada' - referencia a drogas
Xclusivo (Remix)	Me tienes loco con ese cuerpo' ; 'Ese culo pone a las demá' insegura' ; 'Mezcla lo sensual' ; 'No le hablen de embarazo ni de prueba' ; 'El alcohol hizo que a la amiga quiera besar' ; ''Te gusta cuando te lo hago'
Esclava (remix)	Hoy serás mi esclava en el cuarto de un motel' ; 'Quédate en pantys' ; 'Te gusta, en la cama, como te maltrato' ; 'La máquina que vibra' ; 'Que, nada más con tocarla, yo tengo el poder de lograr que la piel se le erice'
Halo	Y eso que yo no soy malo, pero tengo un palo. Que, si abre la boca, te va a dejar un regalo' ; 'Ese culo está grande, eso no es humano. Pa' agarrarlo bien tengo que usar las dos mano' ; 'Ese pussy está clean, me tiene comiendo sano'

Tabla 2.1

Canción	Artista	Realidad	Predicción	CONCLUSIÓN
Mr Moondial	Quevedo	SÍ EXPLÍCITA	CORRECTO	VERDADERO POSITIVO
Titi Me Preguntó	Bad Bunny	SÍ EXPLÍCITA	CORRECTO	VERDADERO POSITIVO
SIn Pijama	Becky G	SÍ EXPLÍCITA	CORRECTO	VERDADERO POSITIVO
Duro de Verdad	Bad Gyal	SÍ EXPLÍCITA	FALSO	FALSO NEGATIVO
Fanatica sensual	Plan B	SÍ EXPLÍCITA	CORRECTO	VERDADERO POSITIVO
Candy	Plan B	SÍ EXPLÍCITA	CORRECTO	VERDADERO POSITIVO
Mi Vecinita	Plan B	SÍ EXPLÍCITA	CORRECTO	VERDADERO POSITIVO
Perro Negro	Bad Bunny	SÍ EXPLÍCITA	CORRECTO	VERDADERO POSITIVO
SI SSABE FERXXO	Blessed, Feid	SÍ EXPLÍCITA	CORRECTO	VERDADERO POSITIVO
LALA	Myke Towers	SÍ EXPLÍCITA	CORRECTO	VERDADERO POSITIVO
DEPORTIVO	Blessed, Anuel AA	SÍ EXPLÍCITA	CORRECTO	VERDADERO POSITIVO
Qué pecao	Manuel Turizo	SÍ EXPLÍCITA	FALSO	FALSO NEGATIVO
Oe Bebé	Blessed, Maluma	SÍ EXPLÍCITA	FALSO	FALSO NEGATIVO
Passoa	JHAYCO, Kapo	SÍ EXPLÍCITA	CORRECTO	VERDADERO POSITIVO
Godiva	Ovy on the Drums	SÍ EXPLÍCITA	CORRECTO	VERDADERO POSITIVO
Amargura	Karol G	NO EXPLÍCITA	CORRECTO	VERDADERO NEGATIVO
Según quien	Maluma, Carín León	NO EXPLÍCITA	CORRECTO	VERDADERO NEGATIVO
Contigo	Karol G	NO EXPLÍCITA	CORRECTO	VERDADERO NEGATIVO
Soltera	Shakira	NO EXPLÍCITA	CORRECTO	VERDADERO NEGATIVO

Orión	Boza, Elena Rose	NO EXPLÍCITA	FALSO	FALSO POSITIVO
Bailando por ahi	Juan Magán	NO EXPLÍCITA	CORRECTO	VERDADERO NEGATIVO
Entre la playa ella y yo		NO EXPLÍCITA	CORRECTO	VERDADERO NEGATIVO
LIMBO	Daddy Yankee	NO EXPLÍCITA	CORRECTO	VERDADERO NEGATIVO
Te pintaron pajaritos	Yandar y Yostin	NO EXPLÍCITA	CORRECTO	VERDADERO NEGATIVO
Ohnana	Kapo	NO EXPLÍCITA	CORRECTO	VERDADERO NEGATIVO
Imaginate	Dany Ocean	NO EXPLÍCITA	CORRECTO	VERDADERO NEGATIVO
Quiéreme mientras se pueda	Manuel Turizo	NO EXPLÍCITA	CORRECTO	VERDADERO NEGATIVO
NUEVAYoL	Bad Bunny	NO EXPLÍCITA	CORRECTO	VERDADERO NEGATIVO
Capaz	Alleh, Yorghaki	NO EXPLÍCITA	CORRECTO	VERDADERO NEGATIVO
DTMF	Bad Bunny	NO EXPLÍCITA	FALSO	FALSO POSITIVO

Tabla 3.1

Canción	Artista	Realidad	Predicción	CONCLUSIÓN
Frente al Mar	Beéle	SÍ EXPLÍCITA	CORRECTO	VERDADERO POSITIVO
AGRADECIDOS	Kaydy Cain	SÍ EXPLÍCITA	CORRECTO	VERDADERO POSITIVO
Fardos	JC Reyes, De la Guetto	SÍ EXPLÍCITA	CORRECTO	VERDADERO POSITIVO
Tú me calmas	Saiko	SÍ EXPLÍCITA	FALSO	FALSO NEGATIVO
Triste Verano	Eladio Carrión, Anuel AA	SÍ EXPLÍCITA	CORRECTO	VERDADERO POSITIVO
Cuando Sera	Mora, Lunay	SÍ EXPLÍCITA	CORRECTO	VERDADERO POSITIVO
Soy esclavo de tu cuerpo	Yampi, Anuel AA	SÍ EXPLÍCITA	CORRECTO	VERDADERO POSITIVO
Tanta Droga	Eladio Carrión	SÍ EXPLÍCITA	CORRECTO	VERDADERO POSITIVO
Movezz en silencio	Cruz Cafuné	SÍ EXPLÍCITA	FALSO	FALSO NEGATIVO
Sigo enamoraú	Eladio Carrión, Yandel	SÍ EXPLÍCITA	CORRECTO	VERDADERO POSITIVO
No te quieren conmigo	Gaby Music, Lunay, Luar La L	SÍ EXPLÍCITA	CORRECTO	VERDADERO POSITIVO
Luces de neón	Myke Towers	SÍ EXPLÍCITA	CORRECTO	VERDADERO POSITIVO
FERXXO	Feid, ICON	SÍ EXPLÍCITA	CORRECTO	VERDADERO POSITIVO
Tumbado en el jardín	San Serno	SÍ EXPLÍCITA	FALSO	FALSO NEGATIVO
Nube Negra	Yung Dupe	SÍ EXPLÍCITA	FALSO	FALSO NEGATIVO
Ojos verdes	Delaossa	NO EXPLÍCITA	CORRECTO	VERDADERO NEGATIVO
150 canciones	Recycled J, Selecta	NO EXPLÍCITA	CORRECTO	VERDADERO NEGATIVO
Intocable (remix)	Rels B	NO EXPLÍCITA	CORRECTO	VERDADERO NEGATIVO
Dolerme	Rosalía	NO EXPLÍCITA	CORRECTO	VERDADERO NEGATIVO
Ya no te hago falta	San Senra	NO EXPLÍCITA	CORRECTO	VERDADERO NEGATIVO
Felicidades	Delaossa	NO EXPLÍCITA	CORRECTO	VERDADERO NEGATIVO

Angelito	Bad Gyal	NO EXPLÍCITA	CORRECTO	VERDADERO NEGATIVO
Brillos Platino	Almácor	NO EXPLÍCITA	CORRECTO	VERDADERO NEGATIVO
EL RITMO QUE NOS UNE	Ryan Castro	NO EXPLÍCITA	CORRECTO	VERDADERO NEGATIVO
Mírame	Blessd, Ovy on the Drums	NO EXPLÍCITA	CORRECTO	VERDADERO NEGATIVO
GATA	Ralphie Choo	NO EXPLÍCITA	CORRECTO	VERDADERO NEGATIVO
RECHÁZAME	Prince Royce	NO EXPLÍCITA	CORRECTO	VERDADERO NEGATIVO
MORENA	Beéle	NO EXPLÍCITA	CORRECTO	VERDADERO NEGATIVO
TU FEO	Lenny Tavárez, Prince Royce	NO EXPLÍCITA	CORRECTO	VERDADERO NEGATIVO
La Carretera	Prince Royce	NO EXPLÍCITA	CORRECTO	VERDADERO NEGATIVO

Tabla 4.1

CANCIÓN	CLASIFICACIÓN SEGÚN DETECTOR CONTENIDO EXPLICITO Y SEXUAL	CLASIFICACIÓN SEGÚN CHATGPT	ACUERDO O DESACUERDO	Nivel de dificultad
capaz (merenguetón)	NO EXPLÍCITA	NO EXPLÍCITA	ACUERDO	Fácil
BAILE INolVIDABLE	NO EXPLÍCITA	NO EXPLÍCITA	ACUERDO	Medio
Angelito	NO EXPLÍCITA	NO EXPLÍCITA	ACUERDO	Medio
Mr. Moondial	SÍ EXPLÍCITA	SÍ EXPLÍCITA	ACUERDO	Fácil
Imagínate	SÍ EXPLÍCITA	SÍ EXPLÍCITA	ACUERDO	Fácil
Weltita	SÍ EXPLÍCITA	NO EXPLÍCITA	DESACUERDO	Medio
APT.	SÍ EXPLÍCITA	NO EXPLÍCITA	DESACUERDO	Dificil
QUEVASAHACERHOY?	SÍ EXPLÍCITA	SÍ EXPLÍCITA	ACUERDO	Fácil
Parte & Choke	NO EXPLÍCITA	SÍ EXPLÍCITA	DESACUERDO	Dificil
KLOuFRENS	SÍ EXPLÍCITA	NO EXPLÍCITA	DESACUERDO	Dificil
CAFé CON RON	NO EXPLÍCITA	NO EXPLÍCITA	ACUERDO	Fácil
Khé?	NO EXPLÍCITA	NO EXPLÍCITA	ACUERDO	Fácil
HALO	SÍ EXPLÍCITA	SÍ EXPLÍCITA	ACUERDO	Fácil
LUNA	NO EXPLÍCITA	SÍ EXPLÍCITA	DESACUERDO	Medio
Diosa	NO EXPLÍCITA	SÍ EXPLÍCITA	DESACUERDO	Fácil
Estamo bien	SÍ EXPLÍCITA	NO EXPLÍCITA	DESACUERDO	Dificil
Brindemos (Anuel)	SÍ EXPLÍCITA	NO EXPLÍCITA	DESACUERDO	Fácil
Quando bebe	NO EXPLÍCITA	NO EXPLÍCITA	ACUERDO	Fácil

Taki taki	SÍ EXPLÍCITA	NO EXPLÍCITA	DESACUERDO	Fácil
Culpables	NO EXPLÍCITA	SÍ EXPLÍCITA	DESACUERDO	Fácil
Bubalú	NO EXPLÍCITA	SÍ EXPLÍCITA	DESACUERDO	Fácil
Noche de fantasias	SÍ EXPLÍCITA	SÍ EXPLÍCITA	ACUERDO	Fácil
Es un secreto (plan B)	NO EXPLÍCITA	NO EXPLÍCITA	ACUERDO	Fácil
Mi planta (ñengo flow)	SÍ EXPLÍCITA	SÍ EXPLÍCITA	ACUERDO	Fácil
Con calma	SÍ EXPLÍCITA	SÍ EXPLÍCITA	ACUERDO	Fácil
Ella quiere beber	NO EXPLÍCITA	SÍ EXPLÍCITA	DESACUERDO	Medio
Ganas sobran	SÍ EXPLÍCITA	SÍ EXPLÍCITA	ACUERDO	Medio
Soltera (Shakira)	NO EXPLÍCITA	SÍ EXPLÍCITA	DESACUERDO	Medio
Verte ir (dj luyan)	SÍ EXPLÍCITA	SÍ EXPLÍCITA	ACUERDO	Fácil
Lean (super g)	SÍ EXPLÍCITA	SÍ EXPLÍCITA	ACUERDO	Fácil
Asesina remix (britiago)	SÍ EXPLÍCITA	SÍ EXPLÍCITA	ACUERDO	Fácil
Familia (niccky minag y annuel)	NO EXPLÍCITA	SÍ EXPLÍCITA	DESACUERDO	Medio
Dios Bendiga	SÍ EXPLÍCITA	SÍ EXPLÍCITA	ACUERDO	Medio
Tu no amas	NO EXPLÍCITA	SÍ EXPLÍCITA	DESACUERDO	Medio
ODIO	NO EXPLÍCITA	SÍ EXPLÍCITA	DESACUERDO	Medio
Santa María (Bad Gyal)	SÍ EXPLÍCITA	SÍ EXPLÍCITA	ACUERDO	Fácil
Mi error (eladio Carrion)	NO EXPLÍCITA	SÍ EXPLÍCITA	DESACUERDO	Dificil
Dime cuántas veces (Rels B)	NO EXPLÍCITA	SÍ EXPLÍCITA	DESACUERDO	Dificil
Hookah	NO EXPLÍCITA	SÍ EXPLÍCITA	DESACUERDO	Medio

Elixir (Funz y Babby Loud)	SÍ EXPLÍCITA	SÍ EXPLÍCITA	ACUERDO	Dificil
Tú Foto (Ozuna)	NO EXPLÍCITA	NO EXPLÍCITA	ACUERDO	Medio
Nadie (Farruko)	NO EXPLÍCITA	NO EXPLÍCITA	ACUERDO	Fácil
TE necesito (Anuel)	SÍ EXPLÍCITA	NO EXPLÍCITA	DESACUERDO	Medio
Kelejodan	SÍ EXPLÍCITA	SÍ EXPLÍCITA	ACUERDO	Medio
Pa comprarte el universo	SÍ EXPLÍCITA	SÍ EXPLÍCITA	ACUERDO	Medio
Ojalá te vaya bien	NO EXPLÍCITA	NO EXPLÍCITA	ACUERDO	Medio
LA LATINA (Rels B)	SÍ EXPLÍCITA	SÍ EXPLÍCITA	ACUERDO	Fácil
Culpable (DEVA)	SÍ EXPLÍCITA	NO EXPLÍCITA	DESACUERDO	Fácil
MDLR (Morad)	SÍ EXPLÍCITA	SÍ EXPLÍCITA	ACUERDO	Fácil
Whine Up	NO EXPLÍCITA	SÍ EXPLÍCITA	DESACUERDO	Dificil
Soy Bichote	SÍ EXPLÍCITA	SÍ EXPLÍCITA	ACUERDO	Medio

Declaración de Uso de Herramientas de Inteligencia Artificial Generativa en Trabajos Fin de Grado

ADVERTENCIA: Desde la Universidad consideramos que ChatGPT u otras herramientas similares son herramientas muy útiles en la vida académica, aunque su uso queda siempre bajo la responsabilidad del alumno, puesto que las respuestas que proporciona pueden no ser veraces. En este sentido, NO está permitido su uso en la elaboración del Trabajo fin de Grado para generar código porque estas herramientas no son fiables en esa tarea. Aunque el código funcione, no hay garantías de que metodológicamente sea correcto, y es altamente probable que no lo sea.

Por la presente, yo, [Nombre completo del estudiante], estudiante de [nombre del título] de la Universidad Pontificia Comillas al presentar mi Trabajo Fin de Grado titulado "[Título del trabajo]", declaro que he utilizado la herramienta de Inteligencia Artificial Generativa ChatGPT u otras similares de IAG de código sólo en el contexto de las actividades descritas a continuación [el alumno debe mantener solo aquellas en las que se ha usado ChatGPT o similares y borrar el resto. Si no se ha usado ninguna, borrar todas y escribir “no he usado ninguna”]:

1. **Brainstorming de ideas de investigación:** Utilizado para idear y esbozar posibles áreas de investigación.
2. **Crítico:** Para encontrar contra-argumentos a una tesis específica que pretendo defender.
3. **Referencias:** Usado conjuntamente con otras herramientas, como Science, para identificar referencias preliminares que luego he contrastado y validado.
4. **Metodólogo:** Para descubrir métodos aplicables a problemas específicos de investigación.
5. **Interpretador de código:** Para realizar análisis de datos preliminares.
6. **Estudios multidisciplinares:** Para comprender perspectivas de otras comunidades sobre temas de naturaleza multidisciplinar.
7. **Constructor de plantillas:** Para diseñar formatos específicos para secciones del trabajo.
8. **Corrector de estilo literario y de lenguaje:** Para mejorar la calidad lingüística y estilística del texto.
9. **Generador previo de diagramas de flujo y contenido:** Para esbozar diagramas iniciales.
10. **Sintetizador y divulgador de libros complicados:** Para resumir y comprender literatura compleja.
11. **Generador de datos sintéticos de prueba:** Para la creación de conjuntos de datos ficticios.
12. **Generador de problemas de ejemplo:** Para ilustrar conceptos y técnicas.
13. **Revisor:** Para recibir sugerencias sobre cómo mejorar y perfeccionar el trabajo con diferentes niveles de exigencia.
14. **Generador de encuestas:** Para diseñar cuestionarios preliminares.
15. **Traductor:** Para traducir textos de un lenguaje a otro.

Afirmo que toda la información y contenido presentados en este trabajo son producto de mi investigación y esfuerzo individual, excepto donde se ha indicado lo contrario y se han dado los créditos correspondientes (he incluido las referencias adecuadas en el TFG y he explicitado para que se ha usado ChatGPT u otras herramientas similares). Soy consciente de las implicaciones académicas y éticas de presentar un trabajo no original y acepto las consecuencias de cualquier violación a esta declaración.

Fecha: [Fecha]

Firma: _____