Universidad Politécnica
de Madrid

Escuela Técnica Superior de
Ingenieros Informáticos

Grado en Matemáticas e Informática

Trabajo Fin de Grado

Análisis de los Efectos del Consumo Recreativo de Drogas en Publicaciones de Twitter

Autora: Alba Cortés Iglesias

Tutora: Lucía Prieto Santamaría

Cotutor: Alejandro Rodríguez González

Madrid, Junio 2025

Este Trabajo Fin de Grado se ha depositado en la ETSI Informáticos de la Universidad Politécnica de Madrid para su defensa.

Trabajo Fin de Grado

Grado en Matemáticas e Informática

Título: Análisis de los Efectos del Consumo Recreativo de Drogas en Publicaciones de Twitter

2025

Autora: Alba Cortés Iglesias

Tutora:

Lucía Prieto Santamaría

Lenguajes y Sistemas Informáticos e Ingeniería de Software

ETSI Informáticos

Universidad Politécnica de Madrid

Cotutor:

Alejandro Rodríguez González

Lenguajes y Sistemas Informáticos e Ingeniería de Software

ETSI Informáticos

Universidad Politécnica de Madrid

Resumen

Las redes sociales, como Twitter (ahora X), se han convertido en un lugar clave donde la gente habla abiertamente sobre el consumo de drogas recreativas. Este proyecto se centra en el análisis de los efectos del consumo recreativo del éxtasis, el GHB y el 2C-B, a través del estudio de 6.755.394 de tweets, empleando técnicas de procesamiento de lenguaje natural y aprendizaje automático.

Para lograr el objetivo, se aplicó la metodología CRISP-DM, estructurando el trabajo en seis fases. En la comprensión del negocio se revisaron estudios sobre el uso de redes sociales para analizar el consumo de sustancias y la aplicación de técnicas de procesamiento de lenguaje. La comprensión de los datos consistió en estudiar la estructura de los archivos JSON obtenidos desde la API de Twitter y recolectar algunos datos en tablas Excel, lo que permitió definir el esquema de la base de datos. En la fase de preparación se procesaron los datos para insertarlos en la base relacional y se generó el *dataset* final, incluyendo variables como el texto del tweet, métricas o la droga mencionada, entre otras. En la etapa de modelado se aplicaron distintos algoritmos de *Machine Learning* bajo cuatro enfoques, y en la fase de evaluación se analizaron los resultados obtenidos.

Estos resultados demuestran que Twitter es una fuente útil para seguir de cerca el consumo de drogas. Se observó que aproximadamente el 95% de los tweets expresaban experiencias negativas, y para su detección se utilizaron términos MetaMap y expresiones coloquiales (*slang*) clasificados por anotadores según si reflejaban experiencias positivas o negativas relacionadas con el consumo de las sustancias. Además, los modelos de clasificación desarrollados alcanzaron un *F1-Score* superior al 80% en la detección de los efectos, siendo *XGBoost* el algoritmo que ofreció el mejor rendimiento. En definitiva, este estudio sugiere que las redes sociales son una fuente de datos valiosa para analizar percepciones y tendencias en el consumo de drogas, con aplicaciones relevantes en salud pública.

Palabras clave: Drogas recreativas, Twitter, Redes Sociales, Procesamiento de Lenguaje Natural, Aprendizaje Automático, CRISP-DM.

Abstract

Social networks, such as Twitter (now X), have become a key place where people talk openly about recreational drug use. This project focuses on the analysis of the effects of recreational use of ecstasy, GHB and 2C-B, through the study of 6.755.394 tweets, using natural language processing (NLP) and machine learning techniques.

To achieve the objective, the CRISP-DM methodology was applied, structuring the work in six phases. In business understanding, studies on the use of social networks to analyze substance use and the application of language processing techniques were reviewed. Data understanding consisted of studying the structure of the JSON files obtained from the Twitter API and collecting some data in Excel tables, which allowed the database schema to be defined. In the preparation phase, the data were processed to insert them into the relational database and the final dataset was generated, including variables such as the text of the tweet, metrics or the drug mentioned, among others. In the modelling stage, different Machine Learning algorithms were applied under four approaches, and in the evaluation phase, the results obtained were analyzed.

These results demonstrate that Twitter is a useful source for monitoring drug use. It was observed that approximately 95% of the tweets expressed negative experiences, and MetaMap terms and colloquial expressions (slang) classified by annotators according to whether they reflected positive or negative experiences related to substance use were used for their detecting the effects, with XGBoost being the algorithm that provided the best performance. In short, this study suggests that social networks are a valuable data source for analyzing perceptions and trends in drug use, with relevant applications in public health.

Keywords: Recreational Drugs, Twitter, Social Networks, Natural Language Processing (NLP), Machine Learning (ML), CRISP-DM.

Tabla de contenidos

1 Introducción 1
1.1 Motivación del estudio 1
1.2 Contexto y antecedentes 1
1.3 Visión general 1
1.4 Objetivos 2
1.5 Plan de trabajo y cronograma 2
1.6 Estructura de la memoria 3
2 Estado del arte 4
2.1 Las redes sociales como fuente de datos 4
2.2 Procesamiento de lenguaje natural y aprendizaje automático 4
2.3 Metodologías estructuradas para el análisis de datos 5
2.4 Perspectivas y consideraciones finales 5
3 Metodología 6
3.1 Metodología CRISP-DM 6
3.2 Aplicación de CRISP-DM al caso de estudio 8
3.3 Herramientas utilizadas 10
3.4 Estructura del repositorio del proyecto 12
3.5 Entorno de ejecución 12
4 Comprensión de los datos 14
4.1 Datos sin procesar en archivos JSON 14
4.2 Datos en archivos Excel 14
5 Preparación de los datos 18
5.1 Esquema del modelo Entidad-Relación 18
5.2 Implementación de la base de datos 19
6 Análisis exploratorio 22
6.1 Tweets 23
6.2 Temporalidad 24
6.3 Geolocalización 25
6.4 Usuarios 27
6.5 Referencias 28
6.6 Engagement 28
6.7 Slang 29
6.8 Metamap 30
7 Modelado 32
7.1 Preparación del dataset 32
7.2 Ingeniería de características y preprocesado 34
7.3 Entrenamiento y evaluación de modelos de clasificación 36

8 Discusión y conclusiones 45
8.1 Conclusiones 45
8.2 Limitaciones 45
8.3 Líneas de trabajo futuras 46
9 Análisis de impacto 47
9.1 Impacto personal 47
9.2 Impacto social 47
9.3 Impacto económico y empresarial 47
9.4 Impacto medioambiental 48
9.5 Impacto cultural 48
9.6 Referencia a los Objetivos de Desarrollo Sostenible de la Agenda 2030 48
9.7 Decisiones basadas en la consideración del impacto 48
10 Bibliografía 49
Anexo 52

Índice de Figuras

Figura 1. Diagrama de Gantt: fases y temporalización del presente TFG..... 3
Figura 2. Ciclo de vida de la metodología CRISP-DM. Reproducida de [10]. 6
Figura 3. Tareas y resultados de la metodología CRISP-DM. Reproducida de [10]. 8
Figura 4. Esquema de las tablas de la base de datos en MySQL. 19
Figura 5. Diagrama de flujo de procesamiento de tweets desde archivos JSON. 22
Figura 6. Distribución de tweets por droga..... 23
Figura 7. Tweet que menciona la palabra "MDMA" 24
Figura 8. Distribución de tweets por año..... 24
Figura 9. Distribución de tweets por mes. 25
Figura 10. Mapa de calor de la distribución global de los tweets..... 26
Figura 11. Tweet publicado por el usuario verificado MAPS 27
Figura 12. Distribución y promedio de interacciones por droga. 28
Figura 13. Matriz de correlación de las interacciones. 29
Figura 14. Distribución de TUIs: colores y frecuencia en tweets. 31
Figura 15. Diagrama de flujo del proceso de selección de tweets. 34
Figura 16. Comparación de los modelos en test sin técnicas de balanceo (F1-Score y AUPRC). 40
Figura 17. Comparación de F1-Score en Cost-Sensitive. 41
Figura 18. Comparación del F1-Score con SMOTE en CV y test 42
Figura 19. Desempeño de los modelos en test con SMOTE In-CV: F1-Score y AUPRC. 43

Índice de Tablas

Tabla 1. Identificación y descripción de las drogas. 15
Tabla 2. Ejemplos de keywords asociadas a las drogas principales..... 15
Tabla 3. Conceptos y anotaciones de slang por cada experto..... 16
Tabla 4. Términos extraídos con MetaMap y anotaciones asociadas. 17
Tabla 5. Tipos semánticos asociados a los términos MetaMap. 17
Tabla 6. Número de tweets por país. 26
Tabla 7. Rendimiento de modelos sin oversampling (conjunto de prueba)..... 39
Tabla 8. Rendimiento de modelos Cost-Sensitive (conjunto de prueba)..... 40
Tabla 9. Rendimiento de modelos con SMOTE Pre-CV (conjunto de prueba). 41
Tabla 10. Rendimiento de modelos con SMOTE In-CV (conjunto de prueba). 42

Lista de Acrónimos

ACC	*Accuracy*
ADASYN	*Adaptative Synthetic Sampling*
AUROC	*Area Under the Receiver Operating Characteristic Curve*
AUPRC	*Area Under the Precision-Recall Curve*
BERT	*Bidirectional Encoder Representations from Transformers*
CRISP-DM	*Cross-Industry Standard Process for Data Mining*
CNN	*Convolutional Neural Networks*
CTB	Centro de Tecnología Biomédica
CUI	*Concept Unique Identifier*
CV	*Cross-Validation*
DT	*Decision Tree*
EDA	*Exploratory Data Analysis*
ER	Entidad-Relación
ETL	*Extract, Transform, Load*
FK	*Foreign Key*
FN	*False Negative*
FP	*False Positive*
IDE	*Integrated Development Environment*
kNN	*k-Nearest Neighbors*
LLM	*Large Language Model*
LR	*Logistic Regression*
ML	*Machine Learning*
MLP	*Multi-Layer Perceptron*
NLP	*Natural Language Processing*
ORM	*Object-Relational Mapping*
PK	*Primary Key*
PREC	*Precision*
REC	*Recall*
RF	*Random Forest*
SMOTE	*Synthetic Minority Over-sampling Technique*
SQL	*Structured Query Language*
SVM	*Suppor Vector Machine*
LDA	*Latent Dirichlet Allocation*
TFG	Trabajo de Fin de Grado
TN	*True Negative*
TP	*True Positive*
TUI	*Type Unique Identifier*
UMLS	*Unified Medical Language System*
UPM	Universidad Politécnica de Madrid
WP	*Work Package*
XGBoost	*Extreme Gradient Boosting*

1 Introducción

1.1 Motivación del estudio

El consumo recreativo de drogas es un fenómeno social muy relevante, con importantes consecuencias tanto para la salud pública como para el bienestar de las personas. Hasta ahora, el estudio de este tema se ha basado principalmente en encuestas, investigaciones clínicas y análisis de datos sanitarios. No obstante, el auge de las redes sociales ha abierto nuevas vías de análisis, destacando particularmente Twitter por permitir la publicación rápida de mensajes breves, lo que le convierte en una fuente de información única para observar cómo se habla públicamente sobre este asunto en tiempo real.

El trabajo propone aprovechar estos datos para complementar el conocimiento existente y obtener una visión más cerca a la experiencia de los usuarios. En concreto, se analiza si los tweets reflejan efectos positivos o negativos tras el consumo de drogas recreativas, lo que puede revelar patrones difíciles de detectar en estudios convencionales y aportar información útil para diseñar mejores estrategias de salud pública.

1.2 Contexto y antecedentes

Este Trabajo de Fin de Grado (TFG) se sitúa en la intersección entre la ciencia de datos, el procesamiento del lenguaje natural y el estudio del comportamiento social en entornos digitales. En concreto, forma parte de una línea de investigación del *Medical Data Analytics Laboratory* (MEDAL) del Centro de Tecnología Biomédica (CTB) de la Universidad Politécnica de Madrid (UPM). Cabe señalar que este proyecto no comienza desde cero, ya que MEDAL con la colaboración de investigadores del Departamento de Psicología Clínica y Experimental de la Universidad de Huelva, que aportan una visión experta en el área del consumo de drogas desde el punto de vista de la psicología, ya habían desarrollado previamente una recopilación de tweets y un análisis preliminar sobre el consumo recreativo de drogas. Por ello, este TFG se plantea como una continuación y un desarrollo más profundo de ese trabajo inicial.

1.3 Visión general

Este proyecto se enfoca en la aplicación de la metodología *Cross Industry Standard Process for Data Mining* CRISP-DM para analizar datos de Twitter relacionados con el consumo recreativo de drogas, con el propósito de identificar si los usuarios reportan efectos positivos o negativos tras su consumo. Cabe destacar que el estudio se limita específicamente al análisis de tres drogas: éxtasis, GHB y 2C-B. Se partirá de un conjunto de datos ya recopilado y almacenado en formato JSON, que será procesado mediante un procedimiento estructurado que incluye la ingesta, limpieza y almacenamiento en una base de datos relacional. A continuación, se realizará un análisis exploratorio detallado para entender las características y tendencias de las publicaciones y, la etapa central del proyecto se centrará en la identificación y clasificación de los efectos mencionados, explorando la aplicación de técnicas de *Natural Language Processing* (NLP) y *Machine Learning* (ML).

1.4 Objetivos

El **objetivo principal** es analizar si los usuarios de Twitter reportan efectos positivos o negativos tras el consumo recreativo de drogas. Para ello, se aplicará la metodología CRISP-DM como marco estructurado de análisis de datos y, además, se propondrá un modelo de clasificación que categorice los tweets según los efectos reportados.

Para alcanzar el objetivo general, se plantean los siguientes **objetivos específicos**:

- Aplicar rigurosamente la metodología CRISP-DM para guiar todas las fases del análisis de datos extraídos de Twitter.
- Diseñar y ejecutar un proceso sólido de Extracción, Transformación y Carga (*Extract, Transform and Load* (ETL), por sus siglas en inglés) que permita limpiar, transformar y almacenar los datos en una base de datos relacional.
- Realizar un análisis descriptivo detallado para explorar las características, tendencias y patrones iniciales relacionados con el consumo recreativo de drogas en redes sociales.
- Investigar y aplicar técnicas de NLP para identificar y extraer los efectos mencionados por los usuarios en sus publicaciones.
- Clasificar los efectos identificados en dos categorías principales: positivos y negativos, y analizar posibles patrones o correlaciones en los datos.
- Aplicar técnicas de ML para automatizar la clasificación o predicción de estos efectos, en función de la calidad y cantidad de los datos disponibles.

1.5 Plan de trabajo y cronograma

La consecución de estos objetivos se abordará a través de un plan de trabajo estructurado en varios *Work Packages* (WP), con una dedicación total estimada de 324 horas:

- **WP0: Programa de orientación laboral.** Se ejecutará de forma paralela al proyecto para complementar la formación y preparación profesional.
- **WP1: Investigación preliminar:** Lectura de TFG previo de referencia, estudios relevantes y comprensión de la metodología CRISP-DM.
- **WP2: Preparación de los datos:** Modelado de un esquema de base de datos relacional, limpieza, transformación y carga de los datos en la base de datos diseñada.
- **WP3: Análisis exploratorio de datos.** Cálculo de estadísticas descriptivas, visualización de variables, identificación de tendencias y entendimiento para posterior modelado.
- **WP4: Modelado.** Aplicación de técnicas de aprendizaje automático, evaluación de modelos y selección de la mejor estrategia analítica.
- **WP5: Interpretación y evaluación.** Interpretación de los hallazgos, comparación con trabajos previos y extracción de conclusiones clave.
- **WP6: Redacción de la memoria final.** Elaboración del documento del TFG, revisión, mejora del contenido y preparación de la presentación para la defensa.

1.5.1 Diagrama de Gantt

A continuación, en la **Figura 1** se presenta el Diagrama de Gantt realizado con Project Libre, que esquematiza la temporalización del proyecto.

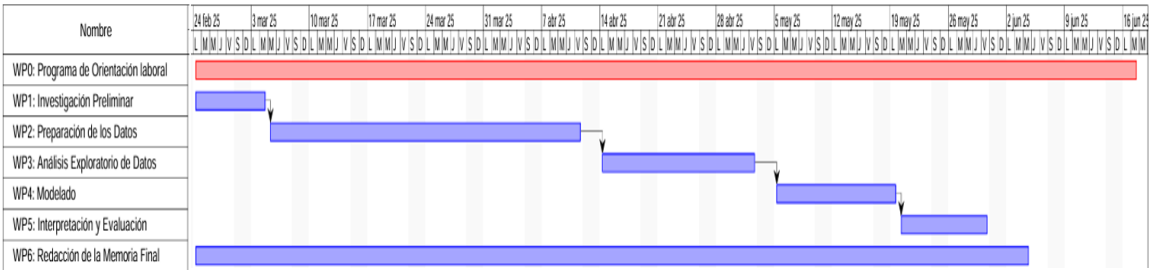


Figura 1. Diagrama de Gantt: fases y temporalización del presente TFG.

1.6 Estructura de la memoria

Esta memoria está organizada para presentar de manera clara y ordenada el desarrollo de un proceso completo de análisis de datos, estructurado según la metodología CRISP-DM. El contenido se divide en los siguientes capítulos:

- **Capítulo 1: Introducción.** Presenta el contexto del estudio, la motivación, una visión general del proyecto, los objetivos, el plan de trabajo, el cronograma y la estructura de la memoria.
- **Capítulo 2: Estado del arte.** Revisa la literatura relevante sobre análisis de redes sociales, procesamiento de lenguaje natural aplicado a Twitter y estudios del consumo de drogas mediante datos digitales.
- **Capítulo 3: Metodología.** Describe la metodología CRISP-DM aplicada al caso de estudio, las herramientas utilizadas y la estructura del repositorio del proyecto.
- **Capítulo 4: Comprensión de los datos.** Expone las características de los datos originales, detallando las tablas complementarias.
- **Capítulo 5: Preparación de los datos.** Detalla el proceso de Extracción, Transformación y Carga, el diseño del modelo entidad-relación y la construcción de la base de datos relacional.
- **Capítulo 6: Análisis exploratorio.** Presenta un análisis descriptivo de los datos, incluyendo temporalidad, geolocalización, usuarios, engagement, uso de slang y conceptos biomédicos extraídos con MetaMap.
- **Capítulo 7: Modelado.** Explora la aplicación de técnicas de aprendizaje automático para clasificar los efectos reportados, detallando enfoques, evaluación y resultados.
- **Capítulo 8: Discusión y conclusiones.** Analiza los hallazgos principales, sus implicaciones, las limitaciones del estudio y las conclusiones generales.
- **Capítulo 9: Análisis del impacto.** Evalúa el impacto potencial del trabajo a nivel académico, social y en el ámbito de la salud pública.
- **Capítulo 10: Bibliografía.** Reúne todas las fuentes y referencias utilizadas.
- **Anexos.** Contiene material complementario como esquemas y figuras adicionales del análisis.

2 Estado del arte

A continuación, se abordan tres aspectos fundamentales para el desarrollo del proyecto: el uso de las redes sociales como fuente de información, cuáles son las técnicas de procesamiento de lenguaje natural aplicadas al análisis de textos y de *machine learning* para entrenar modelos de clasificación, y además el uso de metodologías estructuradas, como CRISP-DM, para organizar y guiar el análisis de los datos.

2.1 Las redes sociales como fuente de datos

Las redes sociales, especialmente Twitter, se han convertido en un recurso clave para analizar cómo se percibe y cuáles son los efectos del consumo recreativo de drogas. Estas plataformas reflejan actitudes y comportamientos en tiempo real, ofreciendo una ventana única a cómo lo usuarios discuten el uso de sustancias, sus motivaciones y las influencias sociales, desde la normalización hasta los riesgos [1], [2].

La accesibilidad de los datos públicos de Twitter permite un análisis profundo de la percepción del consumo de drogas. Las investigaciones han demostrado que la exposición a contenido positivo sobre sustancias en estas plataformas puede influir significativamente en las actitudes, especialmente entre adolescentes, aumentando la probabilidad de consumo de alcohol, marihuana o cocaína [1], [2], [3]. Además, se ha observado que muchos usuarios publican bajo los efectos de las drogas, lo que a menudo lleva consecuencias sociales y arrepentimiento, evidenciando los riesgos asociados [4]. Por ejemplo, un análisis encontró que el 34,4% de los asistentes a eventos de música electrónica en Nueva York publicaron en redes sociales bajo los efectos de las drogas, y 21,4% lamentó haberlo hecho [4].

Para la recolección de datos, Twitter es una fuente muy útil por su capacidad de ofrecer información en tiempo real a través de su API [2]. Sin embargo, trabajar con estos datos presenta desafíos inherentes al lenguaje informal y a la naturaleza ruidosa de la información, lo que exige un preprocesamiento cuidadoso [5]. Para superar estos obstáculos, se han creado enfoques metódicos que no sólo evalúan la fiabilidad de los datos, sino que también emplean estructuras de conocimiento para identificar y ampliar la búsqueda de temas relevantes en el ámbito del abuso de sustancias [6]. Adicionalmente, el análisis de los niveles de *engagement* de los usuarios puede ofrecer una comprensión más profunda de sus interacciones y percepciones sobre el tema [7].

2.2 Procesamiento de lenguaje natural y aprendizaje automático

El procesamiento de grandes volúmenes de texto de redes sociales se apoya fuertemente en técnicas de NLP. Las aproximaciones iniciales suelen incluir la búsqueda de palabras clave y el análisis de sentimiento (también conocido como minería de opinión), que permite clasificar los tweets como positivos, negativos o neutrales basándose en su polaridad emocional [2], [7], [8].

Para lograr mayor precisión y profundidad, se aplican modelos avanzados de aprendizaje automático y *deep learning*. Se han utilizado redes neuronales convolucionales (*Convolutional Neural Networks* (CNN), por sus siglas en inglés)

combinadas con técnicas de tokenización, alcanzando precisiones superiores al 80% en la detección de drogas [5]. Otros algoritmos de aprendizaje automático ampliamente empleados para la clasificación de sentimiento incluyen *Random Forest*, *Multilayer Perceptron* (MLP), *Suppor Vector Machine* (SVM), C5.0, *Logistic Regression*, *Naive Bayes* y *k-Nearest Neighbors* (kNN), demostrando su eficacia en el análisis de texto [7], [8].

Un reto común en este tipo de análisis es el manejo de conjuntos de datos desbalanceados, donde la representación de una clase (por ejemplo, tweets negativos) es significativamente menor que la de otra. Con el fin de disminuir influencias no deseadas en el análisis, se implementan técnicas de muestreo, como el submuestreo (*under-sampling*) y el sobremuestreo (*over-sampling*), incluyendo métodos como SMOTE y ADASYN, que han probado mejorar la robustez y precisión de los modelos [8].

Además de la clasificación, herramientas como la codificación de palabras, conocidos como métodos de *embedding*, y el modelado de temas con *Latent Dirichlet Allocation (LDA)* son cruciales para identificar patrones semánticos y temáticos subyacentes en las conversaciones, manejando la complejidad y la ambigüedad del lenguaje de redes sociales [5], [6]. Incluso se ha explorado el uso de herramientas comerciales de NLP como *IBM Watson*, *Google Cloud Natural Language* y *Meaning Cloud* para la anotación y clasificación automatizada de sentimientos [8].

2.3 Metodologías estructuradas para el análisis de datos

Aunque no siempre se menciona directamente el uso de CRISP-DM, varios estudios aplican principios similares, como la recopilación, limpieza y análisis estructurado de datos. Por ejemplo, un trabajo desarrolló un modelo de aprendizaje automático para detectar daños relacionados con drogas, siguiendo etapas propias de CRISP-DM, como la preparación de datos y el modelado [9].

Un ejemplo claro de un enfoque sistemático es el de "*Social Media Text Mining Framework for Drug Abuse*" propuesto en la literatura [6]. Este marco, diseñado para identificar el abuso de drogas en tiempo real, utiliza un proceso sistemático que incluye la detección de temas, la recolección de datos, la preparación y evaluación de la calidad de los datos, y finalmente, la selección de enfoques de análisis. Este tipo de metodologías es crucial para gestionar eficazmente los datos complejos y ruidosos de las redes sociales, asegurando resultados estructurados y fiables [6].

2.4 Perspectivas y consideraciones finales

En resumen, las redes sociales son una herramienta poderosa para analizar el consumo recreativo de drogas. Los datos que se obtienen de estas plataformas, cuando son procesados con técnicas de NLP y aprendizaje automático, son capaces de revelar patrones de consumo y efectos sociales. Los métodos estructurados, como los que siguen el modelo CRISP-DM, permiten gestionar de forma eficaz los datos complejos. Además, los resultados de estas investigaciones no solo destacan la importancia de tener en cuenta los posibles riesgos asociados al contenido sobre drogas online, sino que también abren la puesta a identificar patrones positivos y oportunidades de intervención en salud pública.

3 Metodología

3.1 Metodología CRISP-DM

La metodología *Cross Industry Standard Process for Data Mining* (CRISP-DM) se ha consolidado como uno de los marcos más reconocidos en el campo de la ciencia de datos y la minería de datos. Desde su publicación en 1999, se ha destacado por ser su enfoque iterativo, su flexibilidad y su orientación clara hacia los objetivos del proyecto [10]. Esta metodología nació como una respuesta a la necesidad de organizar los procesos de análisis de datos en múltiples industrias, proporcionando una estructura clara para la planificación, ejecución y evaluación de proyectos complejos.

Una de las características más destacadas de CRISP-DM es su naturaleza no lineal, a diferencia de enfoques más rígidos como el modelo en cascada. Sus seis fases no están destinadas a seguir una secuencia fija, lo que permite a los equipos avanzar o retroceder entre etapas según lo requieran los resultados obtenidos en el proceso [10]. Esta capacidad de adaptación hace que CRISP-DM sea ideal para adaptarse a los retos y descubrimientos que aparecen durante el desarrollo del proyecto. Además, su estructura flexible es compatible con metodologías ágiles, ya que facilita un progreso iterativo y la incorporación de ajustes conforme aparecen nuevas necesidades o conocimientos. Su capacidad de adaptación es una de las razones por las que CRISP-DM sigue siendo una herramienta fundamental para estructurar proyectos de ciencia de datos [11].

La **Figura 2** muestra el proceso CRISP-DM y cómo sus fases están interconectadas de manera iterativa, permitiendo flexibilidad en el desarrollo de proyectos de análisis.

Figura 2. Ciclo de vida de la metodología CRISP-DM. Reproducida de [10].

1. Comprensión del negocio

La primera fase del proceso CRISP-DM busca entender los objetivos de la empresa y establece las bases del proyecto. Es esencial para asegurar que el análisis de datos responda a las necesidades reales del negocio. Se identifican las metas principales y las preguntas que se desea resolver, definiendo criterios de éxito claros y medibles, como por ejemplo aumentar ventas en un 10%. También se realiza una evaluación de la situación actual: se recopila información sobre los recursos disponibles (personal, datos, herramientas), se identifican restricciones (tiempo, calidad, seguridad), y se consideran riesgos potenciales. A partir de esta evaluación se definen los objetivos de la minería de datos en términos técnicos (por ejemplo, clasificar clientes o predecir ventas) y se elabora un plan detallado de proyecto que incluye tareas, cronograma, recursos y posibles iteraciones.

Resultados: Declaración de objetivos, evaluación de recursos y riesgos, objetivos técnicos y un plan claro.

2. Comprensión de los Datos

Esta fase se centra en recolectar y explorar los datos disponibles. Es clave para conocer la información que se utilizará posteriormente en el análisis. Se identifican y extraen las fuentes de datos (bases de datos, archivos), documentando los problemas detectados y las soluciones aplicadas. A continuación, se describe el contenido de los datos (tipo de archivo, tamaño, atributos) y se realiza una exploración preliminar utilizando visualizaciones o estadísticas básicas que ayuden a detectar patrones, tendencias o anomalías. Por último, se verifica la calidad de los datos para asegurar que sean completos, consistentes y correctos, y se proponen estrategias para corregir errores como valores faltantes o inconsistencias.

Resultados: Informes sobre la recopilación, descripción, exploración y calidad de los datos [12].

3. Preparación de los Datos

En esta etapa se transforman los datos en bruto en un conjunto listo para el análisis. Generalmente, es la fase más extensa del proceso, pudiendo requerir hasta el 80% del tiempo total del proyecto [12]. Se seleccionan las variables relevantes, se corrigen errores y valores faltantes, y se crean nuevos atributos si es necesario. También se realiza la integración de datos prevenientes de distintas fuentes y se ajusta su formato a las necesidades de las herramientas que se utilizarán en las siguientes fases.

Resultados: Conjunto de datos final limpio y listo.

4. Modelado

En esta fase se aplican técnicas de análisis, como modelos estadísticos o de aprendizaje automático, con el objetivo de resolver el problema definido. Se escoge la técnica de modelado más adecuada (por ejemplo, clasificación o regresión) y se diseña un plan de prueba, dividiendo los datos en conjuntos de entrenamiento y prueba. A partir de ahí, e construyen los modelos y se ajustan sus parámetros. Finalmente, se evalúan con métricas que permitan medir su rendimiento y adecuación al problema.

Resultados: Modelos construidos y evaluados [10].

5. Evaluación

Una vez construidos los modelos, es necesario revisar cuidadosamente sus resultados para confirmar que cumplen con los objetivos de negocio. Se verifica si el modelo responde a las preguntas iniciales, se revisa el proceso seguido para detectar posibles errores o desviaciones, y se decide si es necesario repetir alguna fase o si e proyecto está listo para avanzar hacia su implementación.

Resultados: Evaluación final y plan de acción [11].

6. Despliegue

La fase final consiste en la implementación de los resultados del análisis. Dependiendo del caso, esto puede implicar la integración de un modelo predictivo en un sistema operativo, la entrega de un informe final o la preparación de visualizaciones que sirvan como herramienta de apoyo en la toma de decisiones. También se planifica el monitoreo del sistema para garantizar su rendimiento en el tiempo, y se documenta el proceso para extraer aprendizajes útiles para el futuro.

Resultados: Modelo implementado, informe y recomendaciones [10].

Una vez descritas en detalle las seis fases del modelo CRISP-DM, a continuación, se presenta un resumen visual en la **Figura 3** que recoge las principales actividades asociadas a cada una de ellas. Esta representación facilita la comprensión global del proceso y permite tener una guía clara de los pasos a seguir en un proyecto de minería de datos.

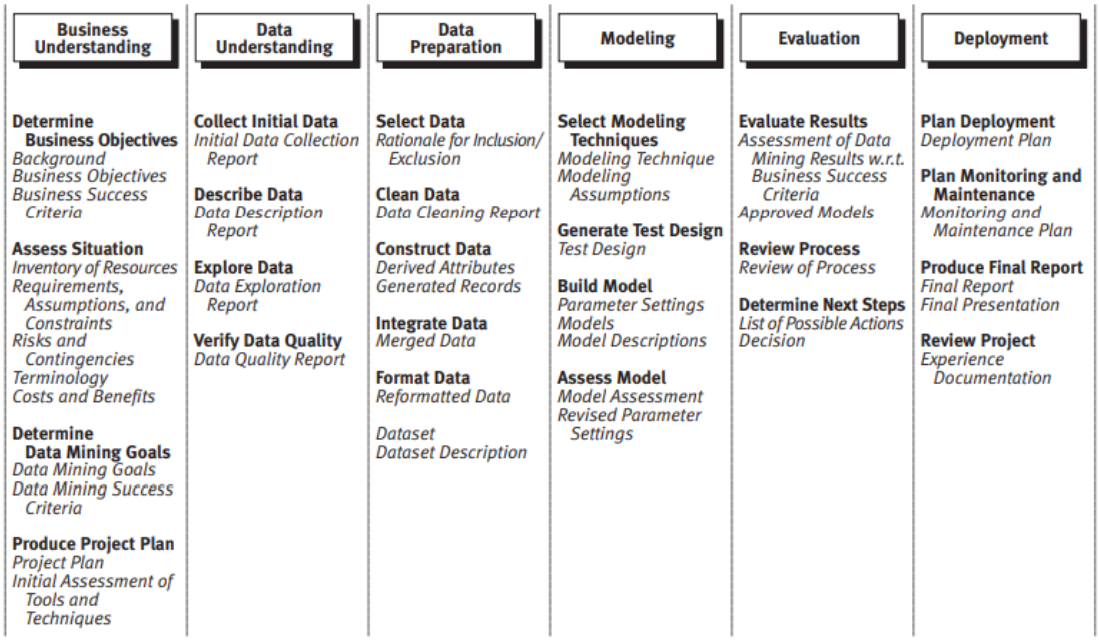


Figura 3. Tareas y resultados de la metodología CRISP-DM. Reproducida de [10].

3.2 Aplicación de CRISP-DM al caso de estudio

La metodología CRISP-DM proporciona un marco estructurado y flexible para proyectos de minería de datos. A continuación, se describe su aplicación al análisis de los efectos del consumo recreativo de drogas en publicaciones de Twitter, siguiendo las fases del proceso.

1. Comprensión del Negocio (*Business Understanding*)

Esta fase inicial fue clave para definir los objetivos del proyecto. El objetivo principal consistió en analizar si los usuarios de Twitter reportan efectos positivos o negativos del consumo recreativo de drogas, complementando estudios tradicionales con datos en tiempo real de redes sociales.

Con el fin de contextualizar el trabajo, se realizó una revisión de estudios previos sobre el uso de redes sociales en investigaciones sobre drogas. Además, se partió de un Trabajo de Fin de Máster (TFM) previo del CTB, aprovechando una base de datos existente [13].

Siguiendo la metodología CRISP-DM, se definieron objetivos técnicos como el diseño del proceso ETL, análisis exploratorio y aplicación de modelos de ML. El proyecto se estructuró en paquetes de trabajo (WPs) y se planificó mediante un diagrama de Gantt, con una dedicación estimada de 324 horas.

2. Comprensión de los Datos (*Data Understanding*)

Se recopilaron aproximadamente 7 millones de tweets en formato JSON (5,71 GB), obtenidos entre 2010 y 2019 a través de la API de Twitter utilizando palabras clave relacionadas con el consumo recreativo de drogas.

Para complementar el conjunto de datos, se desarrollaron hojas de cálculo auxiliares que permitieron organizar información clave como términos de jerga, lista de palabras clave, así como conceptos biomédicos. Este enfoque facilitó una interpretación más precisa del lenguaje presente en los tweets.

Durante la verificación de calidad, se identificaron campos incompletos o nulos, especialmente en la información geográfica y multimedia. Esto creó la necesidad de aplicar un proceso riguroso de limpieza y preparación de datos antes de realizar el análisis.

3. Preparación de los Datos (*Data Preparation*)

Esta fase tuvo como objetivo transformar los datos brutos en un conjunto limpio, estructurado e integrado, listo para el análisis, siendo una de las etapas más exigentes en tiempo. Se seleccionó la información relevante de los archivos JSON y Excel, priorizando el contenido de los tweets, las métricas de interacción y la terminología asociada al consumo de drogas.

Se diseñó un modelo Entidad-Relación en MySQL Workbench para organizar los datos de forma coherente, facilitando su implementación en una base de datos relacional y permitiendo consultas complejas. La base de datos se creó en MySQL siguiendo buenas prácticas de diseño, con claves y restricciones que garantizaron la integridad referencial.

La carga de datos se realizó en varias etapas mediante scripts de Python, utilizando 'pandas' y SQLAlchemy para los archivos Excel, y un sistema modular para procesar los datos JSON. Se aplicaron técnicas como procesamiento por lotes y control de duplicados con el fin de optimizar el rendimiento.

Además, se integró información semántica externa (como conceptos biomédicos MetaMap), lo que enriqueció el conjunto de datos con conocimiento especializado del dominio, difícil de obtener directamente del lenguaje informal de los tweets. Este enfoque robusto y modular permitió preparar una base sólida para el análisis posterior.

4. **Modelado (*Modeling*)**

En esta fase se aplicaron técnicas de aprendizaje automático para clasificar tweets según los efectos positivos o negativos del consumo de drogas. Se probaron diversos algoritmos, como *XGBoost, Random Forest*, Regresión Logística, entre otros, priorizando el *F1-Score* debido al desbalance de clases.

El conjunto de datos se construyó filtrando tweets con polaridad consistente (positiva o negativa) según término de jerga y conceptos biomédicos, lo que aseguró etiquetas fiables, aunque redujo el volumen de datos. Las características utilizadas incluían texto procesado (vectorizado con Word2Vec), interacción, menciones de drogas y datos del usuario.

Se aplicó una validación cruzada estratificada y anidada, y se optimizaron los modelos con *RandomizedSearchCV*. Además, se exploraron distintas estrategias para manejar el desbalance de clases, como el uso de modelos *cost-sensitive* y técnicas de sobremuestreo como SMOTE.

XGBoost y MLP ofrecieron los mejores resultados, incluso sin técnicas explícitas de balanceo, demostrando una alta capacidad para manejar datos desbalanceados. También se identificó una trampa metodológica al aplicar SMOTE antes de la validación cruzada, lo que llevó a un sobreajuste, subrayando la importancia de aplicar correctamente estas técnicas dentro de cada pliegue.

5. **Evaluación (*Evaluation*)**

La validación del modelo se apoyó no solo en métricas rigurosas, sino también en un análisis exploratorio previo que permitió comprender mejor las características y distribución de los datos. Esta exploración inicial ayudó a identificar qué variables estaban disponibles, su formato, nivel de completitud y distribución general, lo que ayudó a definir las decisiones posteriores en cuanto a selección de atributos y preparación del conjunto de entrenamiento.

Para evaluar el rendimiento de los modelos en la clasificación de efectos del consumo de drogas, se utilizaron métricas robustas como *F1-Score, Recall,* AUROC y AUPRC. Se compararon distintos algoritmos (XGBoost, MLP, RF, etc.) y cuatro estrategias de manejo del desbalance, observando que XGBoost y MLP mostraron un rendimiento sólido incluso sin técnicas de balanceo.

3.3 Herramientas utilizadas

3.3.1 Excel

Se utilizó **Microsoft Excel** como herramienta fundamental para la estructuración y organización inicial de los datos auxiliares relacionados con el análisis semántico de los tweets. En particular, se elaboraron cinco hojas de cálculo que contienen información relevante sobre drogas, palabras clave asociadas, jerga, anotaciones semánticas y clasificaciones biomédicas. Estas hojas se estructuraron siguiendo principios de diseño relacional y facilitaron la carga directa en la base de datos MySQL.

3.3.2 MySQL Workbench

Se empleó MySQL Server 8.0, gestionado por **MySQL Workbench,** para el diseño del esquema de la base de datos que representó de forma estructurada las conexiones entre las diferentes entidades del domino, como drogas, palabras clave, tweets y usuarios. Además, se utilizó en conjunto con MySQL Server 8.0

para la creación y gestión de la base de datos, lo que permitió una transición directa desde el modelo lógico al físico. Esta herramienta resultó esencial para garantizar la integridad referencial y facilitar posteriores consultas durante el análisis.

3.3.3 Python

Para el procesamiento y automatización de los flujos de trabajo, es decir, todo el proceso ETL, se empleó **Python 3.12** con un conjunto de librerías especializadas que permitieron la integración eficiente de los datos. Las principales librerías utilizadas fueron:

- 'pandas==2.2.3': Para la lectura, transformación y manipulación de los datos desde archivos Excel y JSON.
- 'SQLAlchemy==2.0.41': Como herramienta ORM para conectar y transferir datos entre los DataFrames de Pandas y la base de datos MySQL, así como también transferir datos entre dos bases de datos MySQL.
- 'mysql-connector-python==9.2.0' o 'PyMySQL==1.1.1': Como *driver* de conexión a MySQL.
- 'json': Para la carga y procesamiento de archivos en formato JSON que contienen los tweets.
- 'os', 'glob': Para la gestión automatizada de directorios y archivos dentro de los scripts.
- 'matplotlib==3.13.3', 'seaborn==0.13.2', 'folium==0.14.0', 'geopandas==0.14.4', 'wordcloud==1.9.3': Para la generación de visualizaciones durante el análisis exploratorio.
- 'nltk==3.9.1': Para tareas de procesamiento de lenguaje natural (NLP), como tokenización, búsqueda de patrones léxicos y normalización del texto.
- 'gensim==4.3.3': Para tareas relacionadas con el modelado de temas o la representación de palabras.
- 'scikit-learn==1.6.1': Para el desarrollo de modelos de aprendizaje automático, especialmente clasificación.
- 'xgboost==3.0.2': Para la implementación de algoritmos de *gradient boosting*.
- 'openpyxl==3.1.5': Para la lectura y escritura de archivos .xlsx de Excel

3.3.4 Jupyter Notebook

Se creó un entorno interactivo con **Jupyter Notebook 7.4.1** que integra consultas SQL, transformaciones con 'pandas' y generación de gráficas. Este entorno permitió una exploración visual ágil de los datos, permitiendo iteraciones rápidas en el análisis y una comunicación efectiva en los resultados obtenidos.

3.3.5 GitHub

Se utilizó **GitHub** como plataforma de control de versiones para el proyecto, permitiendo un seguimiento claro de los cambios realizados en el código, la colaboración en distintas etapas del desarrollo y la gestión ordenada de los archivos del repositorio. Además, sirvió como respaldo centralizado para la documentación técnica y los scripts principales del análisis. El repositorio con

el código y los recursos utilizados en este trabajo está disponible públicamente en GitHub[1].

3.3.6 Visual Studio Code

Visual Studio Code (VSC) fue el entorno de desarrollo integrado principal utilizado para la escritura y edición de los scripts de Python y los Jupyter Notebook. Su versatilidad permitió una navegación eficiente por el código, el uso de extensiones para la depuración y la integración con el control de versiones de Git, facilitando un seguimiento detallado de los cambios. Además, se utilizó para la gestión de archivos y la ejecución de comandos de terminal, optimizando el flujo de desarrollo de todo el proyecto.

3.4 Estructura del repositorio del proyecto

El repositorio del proyecto fue estructurado siguiendo principios de claridad, modularidad y reproducibilidad. Se organizó en carpetas específicas según la funcionalidad y tipo de contenido, con el objetivo de facilitar tanto la navegación como la reutilización del código. A continuación, se describe su estructura principal:

- `/data/`: Contiene los archivos de entrada (`/raw/`), incluyendo hojas de Excel con datos auxiliares y archivos JSON con los tweets. Además, también contiene el dataset final utilizado para el modelado (`/processed/`).
- `/notebooks/`: Incluye los notebooks Jupyter utilizado durante el análisis exploratorio y el análisis de los resultados del modelado.
- `/src/`: Carpeta que contiene los scripts de Python y SQL para automatizar tareas como carga de datos, transformaciones, inserciones a la base de datos, y todo el proceso de modelado.
- `/results/`: Contiene imágenes generadas o incluidas en la memoria, como diagramas del modelo entidad-relación, gráficas del análisis y diagramas de flujo.
- `README.md`: Documento principal del repositorio, donde se resumen las instrucciones de uso y estructura de este.
- `.gitignore`: Archivo de configuración para excluir archivos temporales o datos pesados (como los JSON de los tweets).
- `requirements.txt`: Define las principales librerías necesarias para el funcionamiento del proyecto.

3.5 Entorno de ejecución

El proyecto ha sido desarrollado en un entorno local con las siguientes especificaciones técnicas:

- **Sistema operativo:** Microsoft Windows 11 Pro (64 bits).
- **Procesador:** AMD Ryzen 7 (5000 Series).
- **Memoria RAM**: 12 GB.
- **Almacenamiento**: SSD de 1 TB.
- **Entorno de desarrollo:** Python con ejecución directa desde el sistema.

[1] https://github.com/albacortes9/tfg-analisis-drogas-twitter

- **Editor de código**: Visual Studio Code.
- **Lenguaje de programación**: Python y SQL.
- **Control de versiones:** Git, con repositorio alojado en GitHub.
- **Gestión de paquetes:** Instalación directa con `pip install requirements.txt`.
- **Interfaz interactiva:** Jupyter Notebook (para análisis exploratorio).
- **Base de datos:** MySQL Server 8.0, gestionada mediante MySQL Workbench.

4 Comprensión de los datos

4.1 Datos sin procesar en archivos JSON

Los datos proporcionados inicialmente se componen de 17 archivos en formato JSON que contienen publicaciones de Twitter recopiladas entre los años 2010 y 2019. Estos tweets fueron recolectados a través de la API de Twitter, utilizando como criterio de filtrado diversas palabras clave asociadas al consumo recreativo de drogas, tales como "MDMA", "GEEB", "tucibi", "XTC", entre otras.

Cada archivo incluye información estructurada con los siguientes campos:

- 'author_id': Datos del autor del tweet (nombre, ID, verificación, métricas de seguidores, imagen de perfil y nombre de usuario).
- 'text': Contenido del tweet.
- 'created_at': Fecha y hora de publicación.
- 'entities': Información sobre menciones dentro del tweet.
- 'public_metrics': Métricas de interacción (*retweets*, respuestas, *likes*, citas).
- 'referenced_tweets': Referencias a otros tweets, indicando si se trata de respuestas o citas.
- 'in_reply_to_user_id': Información sobre el usuario al que se responde en caso de ser un tweet de respuesta.
- 'attachments': Datos sobre archivos multimedia adjuntos en los tweets (*url*, tipo, pixeles, ID).
- 'geo': Información sobre la ubicación desde donde se publicó el tweet (id, tipo, coordenadas, nombre, país).

El volumen total de datos asciende a aproximadamente 7.000.000 tweets, ocupando alrededor de 5,71 GB. Este conjunto de datos contiene tanto tweets originales como *retweets* y respuestas, lo que permite analizar tanto el contenido como su impacto o alcance social.

Durante la exploración preliminar se identificaron algunas limitaciones, como la presencia de campos nulos o incompletos, especialmente en 'geo' y 'attachments'.

Esta estructura de datos resulta especialmente adecuada para el análisis propuesto. El campo 'text' será el principal para la aplicación de técnicas de NLP, mientras que las métricas de interacción permitirán contextualizar la recepción de los tweets por parte de otros usuarios. Asimismo, la información geográfica y temporal podría permitir estudios complementarios sobre la evolución y localización del fenómeno.

Este conjunto de datos ofrece una base sólida para explorar los efectos del consumo recreativo de drogas según lo reportado espontáneamente por los usuarios en redes sociales.

4.2 Datos en archivos Excel

Con el objetivo de enriquecer el análisis de los tweets relacionados con el consumo de drogas, se ha elaborado un conjunto de tablas auxiliares estructuradas en un archivo Excel. Estas tablas recopilan información clave sobre las sustancias a estudiar, los términos asociados a ellas (tanto formales

como coloquiales), y anotaciones semánticas que permitirán interpretar el contenido textual desde una perspectiva contextual, afectiva y conceptual.

Estas estructuras de datos servirán de apoyo en fases posteriores del trabajo, tanto para la identificación de patrones como para la construcción del modelo de análisis. A continuación, se describen las principales tablas creadas y su contenido.

4.2.1 Tabla de drogas

La **Tabla 1** contiene las sustancias psicoactivas que serán objeto del análisis. Cada entrada incluye un identificador único, el nombre común de la droga y una breve descripción de sus efectos o características más relevantes.

Tabla 1. Identificación y descripción de las drogas.

id	name	description
1	ghb	Depresor del sistema nervioso.
2	ecstasy	Psicoestimulante, conocido como MDMA.
3	2c-b	Psicodélico derivado de feniletilaminas.

4.2.2 Tabla de palabras clave

La tabla de palabras clave asocia múltiples términos con cada una de las drogas. Estos términos incluyen sinónimos, nombres comerciales, variantes de escritura y expresiones comúnmente utilizadas en redes sociales para referirse a cada sustancia. La tabla contiene en total 45 palabras clave, de las cuales, 13 se corresponden con la droga GHB, 17 al Éxtasis y 15 para el 2C-B. La vinculación entre palabras clave y drogas se establece mediante el campo 'drug_id', que permite construir una relación de uno a muchos entre las dos tablas. Esta información será clave para la detección de menciones indirectas a drogas en el contenido textual. En la **Tabla 2** se muestran algunas de las *keywords* asociadas a la droga GHB.

Tabla 2. Ejemplos de keywords asociadas a las drogas principales.

id	drug_id	keyword
1	1	ghb
2	1	GEEB
3	1	Georgia Home Boy
4	1	liquid ecstasy
5	1	cherry meth
6	1	easy lay
7	1	blue nitro
8	1	gamma hydroxybutyrate

4.2.3 Tabla de *slang*

Con el fin de capturar el lenguaje coloquial presente en los tweets, se ha creado una tabla con términos de jerga relacionados con el consumo de drogas o sus efectos. Esta tabla contiene un total de 89 términos, cada uno ha sido evaluado

por tres personas expertas que han clasificado su significado dentro del contexto de uso en el tweet: positivo, si el término expresa una experiencia placentera o deseada; negativo, sugiere consecuencias perjudiciales o experiencias no deseadas; o contexto, depende del uso, sin connotación clara. En la **Tabla 3** se muestran 10 de las palabras *slang*.

Tabla 3. Conceptos y anotaciones de slang por cada experto.

id	concept	annotation1	annotation2	annotation3
1	addicted	negativo	negativo	negativo
2	anxiety	negativo	negativo	negativo
3	attack	negativo	negativo	negativo
4	bad mood	negativo	negativo	negativo
5	bad trip	negativo	negativo	negativo
6	baked	contexto	neutro	contexto
7	bipolar	negativo	negativo	negativo
8	blasted	negativo	negativo	negativo
9	blazed	contexto	contexto	contexto
10	blotter paper	contexto	contexto	contexto

Para vincular esta información con el resto de la base de datos, se ha seguido un enfoque similar al utilizado en la tabla 'tweet_keyword'. Específicamente, se ha creado una tabla de relación entre tweet y *slang*, denominada 'tweet_slang', en el cual se registran únicamente aquellas coincidencias conectadas entre el contenido textual del tweet y los términos registrados en la tabla *slang*.

El proceso de identificación se realizó buscando coincidencias exactas o parciales de palabras en los textos de los tweets. Es decir, se contemplaron variaciones morfológicas, por ejemplo, si el término original es "anxiety", también se consideró "anxious" como una forma válida para establecer la relación semántica.

4.2.4 Tabla de MetaMap

Los términos recogidos mediante la herramienta MetaMap se corresponden con conceptos biomédicos y conductuales detectados en textos relacionados con el consumo, que se recogen en el sistema *Unified Medical Language System* (UMLS), una referencia estándar en el ámbito biomédico que permite agrupar conceptos según su naturaleza o función [14].

Cada uno ha sido evaluado por tres anotadores, utilizando la misma escala anterior (positivo, negativo, contexto). La **Tabla 4** muestra 10 de los términos de MetaMap, que serán utilizados para el análisis.

Tabla 4. Términos extraídos con MetaMap y anotaciones asociadas.

term	annotation1	annotation2	annotation3
Sexual activity NOS	positivo	contexto	positivo
Behavior, NOS	contexto	contexto	contexto
Ritual	contexto	contexto	positivo
Sexuality	positivo	positivo	positivo
Cruelty towards animals	negativo	negativo	negativo
Gambling	contexto	negativo	contexto
Alcohol consumption	contexto	negativo	contexto
Crying	negativo	contexto	negativo
Singing	positivo	contexto	positivo

4.2.5 Tabla de tipos semánticos

Además de los términos extraídos mediante MetaMap, se ha creado una tabla adicional que clasifica cada uno de ellos según su tipo semántico. En la **Tabla 5** se muestran 10 de los tipos semánticos. Cada uno está representado por un *Type Unique Identifier* (TUI), una abreviación, una breve descripción y, cuando ha sido posible, una anotación que indica el tipo de implicación del concepto dentro del contexto del tweet: física o psicológica.

Tabla 5. Tipos semánticos asociados a los términos MetaMap.

tui	abbreviation	description	annotation
T019	cgab	Congenital Abnormality	
T020	acab	Acquired Abnormality	
T023	bpoc	Body Part, Organ, or Organ Component	
T037	inpo	Injury or Poisoning	
T038	biof	Biologic Function	
T039	phsf	Physiologic Function	Physical
T040	orgf	Organism Function	Physical
T041	menp	Mental Process	Psychological
T046	patf	Pathologic Function	Physical

5 Preparación de los datos

5.1 Esquema del modelo Entidad-Relación

A partir de los datos explorados y estructurados en las fases previas, se diseñó un modelo de Entidad-Relación (ER) con el objetivo de representar de forma coherente, normalizada y relacional la información recopilada de Twitter junto con los recursos auxiliares utilizados para enriquecer el análisis. Este modelo permite una implementación eficiente en una base de datos relacional y facilita la ejecución de consultas complejas sobre el contenido textual, los usuarios, las métricas de interacción y los aspectos contextuales relacionados con el consumo recreativo de drogas.

El esquema de la base de datos fue elaborado utilizado MySQL Workbench. En la **Figura 4** se muestra visualmente esta estructura final, que refleja tanto las entidades principales extraídas de los archivos JSON como aquellas derivadas de las tablas auxiliares creadas en Excel.

Cabe destacar que en **Anexo I** se presenta un esquema correspondiente a un boceto preliminar, creado con el propósito de identificar los atributos clave, los tipos de datos, las claves primarias y foráneas, así como las relaciones entre las distintas tablas. Este diseño sirvió como punto de partida para la implementación final del modelo relacional.

El modelo incluye entidades como:

- *Tweet*: entidad principal, representa cada publicación, con atributos como texto, autor, fecha de publicación y si contiene contenido multimedia.
- *User*: almacena información del autor del tweet, incluyendo nombre, nombre de usuario, localización y métricas de popularidad (seguidores, seguidos, etc.)
- *Ubication*: recoge datos geográficos asociados a los tweets (país, ciudad, tipo de lugar, coordenadas).
- *Tweet_metrics*: contiene métricas de interacción social como retweets, respuestas, likes y citas.
- *Mention* y *Referenced_tweet*: permiten identificar menciones a otros usuarios y relaciones entre tweets (respuestas o citas).
- *Drug* y *Drug_keyword*: definen las drogas a analizar y los términos (keywords) asociados a ellas, que permiten identificar menciones explícitas e implícitas.
- *Tweet_keyword*: relaciona tweets concretos con las palabras clave detectadas en su contenido.
- *Slang* y *Tweet_slang*: incorporan vocabulario de jerga y su aparición en los tweets, junto con anotaciones semánticas subjetivas.
- *Metamap*, *Semantic_type* y *Tweet_metamap*: recogen conceptos detectados por la herramienta MetaMap, clasificados según el sistema UMLS.

Este modelo sirve como base estructural para el análisis posterior. Así, se facilita la detección de patrones relacionados con el discurso sobre drogas en redes sociales, y se garantiza la consistencia de los datos almacenados.

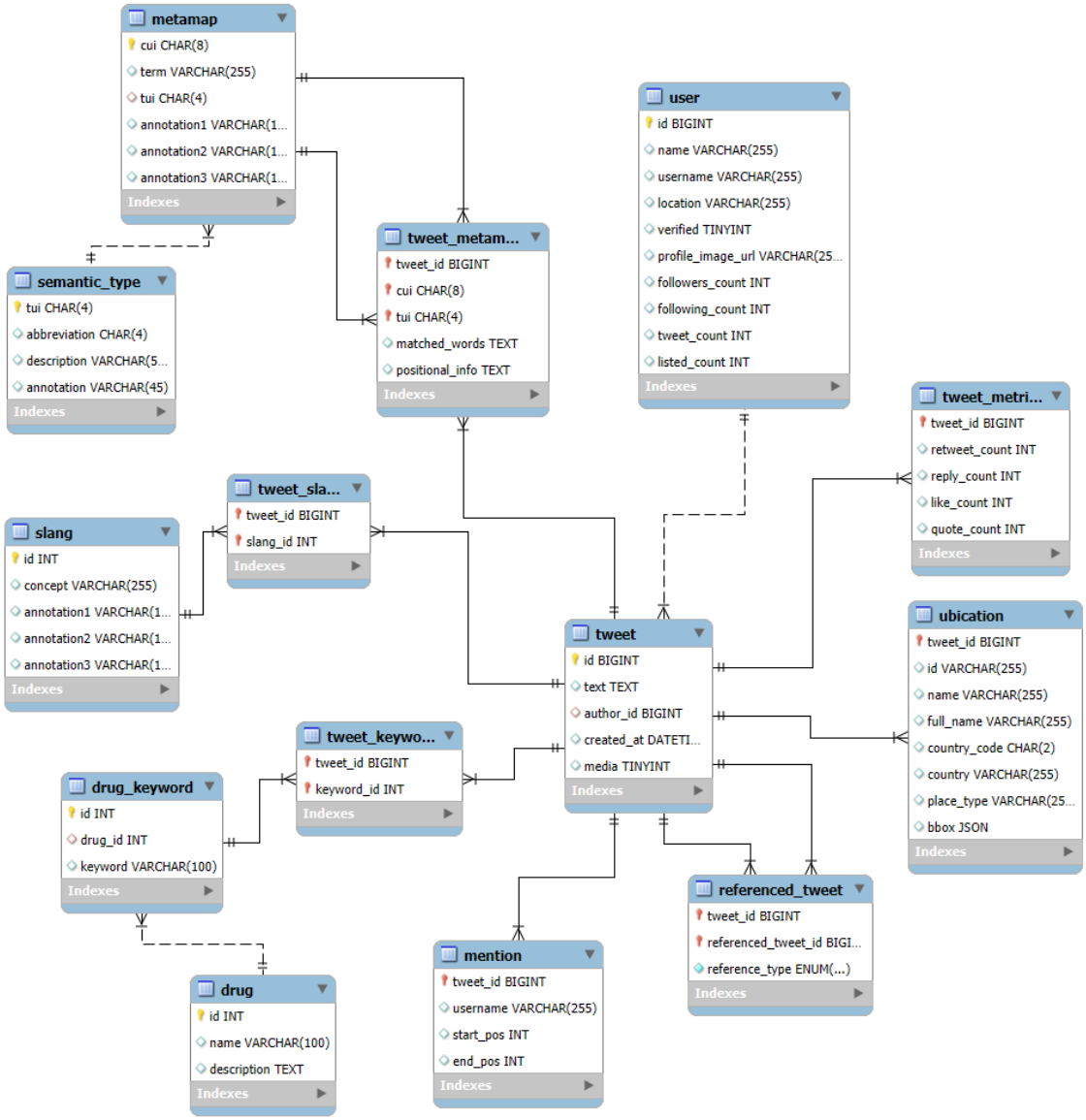


Figura 4. Esquema de las tablas de la base de datos en MySQL.

5.2 Implementación de la base de datos

Para almacenar y gestionar los datos extraídos tanto de archivos JSON como de fuentes complementarias, se diseñó una base de datos relacional utilizando MySQL. Este sistema de almacenamiento permite organizar la información de manera estructurada, garantizar la integridad de los datos y facilitar su consulta posterior para el análisis.

A partir del modelo entidad-relación definido previamente, se generó un script SQL que permite la creación automática de todas las tablas necesarias. Dicho script ha sido desarrollado en MySQL Workbench y se adjunta en el **Anexo II**.

5.2.1 Creación de las tablas

Las tablas fueron creadas siguiendo buenas prácticas de diseño relacional. Algunas de las características principales de la estructura son:

- **Claves primarias (*Primary Key*):** Cada tabla cuenta con una clave primaria compuesta o simple, que garantiza la unicidad de los registros y se define con la restricción `NOT NULL`.
- **Claves foráneas (*Foreign Key*):** Se utilizaron para establecer relaciones entre las tablas, manteniendo la integridad referencial del sistema. Las restricciones aplicadas incluyen:
    - `ON DELETE RESTRICT`: Impide la eliminación de un registro referenciado por otra tabla.
    - `ON UPDATE CASCADE`: Permite que, en caso de actualización de un valor clave, el cambio se propague automáticamente a las tablas que dependan de él.
- **Tipos de datos adecuados:** Se seleccionaron tipos de datos apropiados para cada campo (por ejemplo, `BIGINT` para identificadores de tweets, `TEXT` para textos extensos, `BOOLEAN` para campos binarios), garantizando eficiencia en almacenamiento y consultas.

5.2.2 Inserción de los datos de los Excel

Como parte del proceso de inicialización de la base de datos, se cargaron datos provenientes de un archivo Excel con cinco hojas de cálculo. Cada hoja representaba una entidad estática relacionada con las drogas y su análisis:

- *drug*: Lista de sustancias identificadas.
- *drug_keyword*: Palabras clave asociadas a cada droga.
- *slang*: Términos coloquiales utilizados para referirse a dichas sustancias.
- *metamap*: Identificadores del sistema MetaMap para el procesamiento semántico
- *semantic_type*: Tipos semánticos asociados a cada término médico.

Para su integración en la base de datos, se utilizó Python con la biblioteca Pandas para leer los datos de cada hoja, y SQLAlchemy como *Object-Relational Mapping* (ORM) para realizar la inserción en MySQL. Esta metodología permitió cargar los datos de forma automatizada, segura y reproducible, respetando las restricciones y relaciones definidas en el modelo

Durante la inserción, se emplearon las opciones `if_exists='append'` y `index=False` para añadir los registros sin sobrescribir los existentes y evitar que se insertara el índice generado por 'pandas'.

5.2.3 Inserción de los datos de los JSON

Una vez creadas las tablas y cargados los datos base, se procedió a integrar los datos dinámicos extraídos de Twitter, contenidos en archivos `.json`. Estos archivos almacenan información estructurada sobre los tweets y sus metadatos: usuarios, métricas, menciones, ubicación, referencias y más.

Para procesar estos datos, se desarrolló un sistema modular en Python, compuesto por:

- `utils.py`: Define funciones auxiliares como la conexión a la base de datos (`conectar_bd()`), carga de archivos JSON (`cargar_archvios_json()`) y listado de archivos a procesar (`obtener_archivos_json()`).
- `insertar_*.py`: Cada módulo contiene una función que insertar datos en una tabla específica, transformando previamente los campos si es necesario. Todos siguen una estructura común, utilizando

`executemany()` con lotes (`bacth_size`) para optimizar el rendimiento y `INSERT IGNORE` para evitar conflictos por duplicados.

- `main.py`: Es el script principal del sistema. Recorre todos los archivos `.json` dentro del directorio `data/raw`, los carga uno a uno y ejecuta secuencialmente cada función de inserción.

5.2.4 Inserción de los datos en SQL procesados previamente

Además de los datos provenientes directamente de los archivos JSON, se integraron en el sistema datos procesados previamente y almacenados en otra base de datos independiente. En concreto, se utilizó una tabla llamada 'metamap', la cual contiene información semántica y lingüística extraída del texto de los tweets. Esta tabla incluye, entre otros campos, el identificador del tweet ('tweet_id'), los códigos semánticos *Concept Unique Identifier* (CUI) y TUI, la palabra detectada o coincidente en el texto ('matched_word') y su posición ('positional_info').

Para garantizar la coherencia e integridad de los datos, se realizó una operación de filtrado basada en una sentencia `JOIN` entre la tabla 'metamap' y la tabla 'tweet' de mi base de datos principal. El objetivo fue seleccionar únicamente aquellos registros de 'metamap' cuyo 'tweet_id' tuviera correspondencia con un tweet ya almacenado. De este modo, se evitó insertar información no relevante o desconectada del análisis principal.

Una vez obtenido el conjunto filtrado, se procedió a su inserción en la base de datos 'twitter_analysis'. Al igual que en etapas anteriores, el proceso de carga se diseñó para funcionar por lotes (*batches*), lo que permitió optimizar el rendimiento y evitar problemas de memoria o saturación de la conexión al trabajar con grandes volúmenes de datos.

6 Análisis exploratorio

En esta sección se muestra y analizan visualizaciones generadas a partir de los datos extraídos de Twitter y almacenados en la base de datos relacional implementada previamente. Para ello se ha generado un entorno de análisis en un *Jupyter Notebook* que contiene el código necesario para procesar y representar gráficamente los principales aspectos de los datos.

Como primer paso, se ha construido un diagrama de flujo, **Figura 5**, que resume de forma clara y estructurada las estadísticas más relevantes del conjunto de datos. Este diagrama permite visualizar de manera sintética el volumen total de tweets, su distribución según si mencionan drogas o no, y el desglose posterior por número de sustancias mencionadas (una, dos o más), así como por tipo de droga (éxtasis, GHB y 2C-B). Además, se incorporan datos adicionales como la cantidad de tweets con contenido multimedia o con información de ubicación geográfica, y la proporción de usuarios verificados frente a no verificados.

Esta representación visual no solo facilita una comprensión global del conjunto de datos, sino que también permite identificar patrones relevantes y posibles líneas de investigación. A partir de este análisis exploratorio, se profundizará en otras gráficas más detalladas que examinan aspectos como el lenguaje utilizado (*slang* y conceptos MetaMap), la distribución temporal de los tweets, el comportamiento de los usuarios, etc.

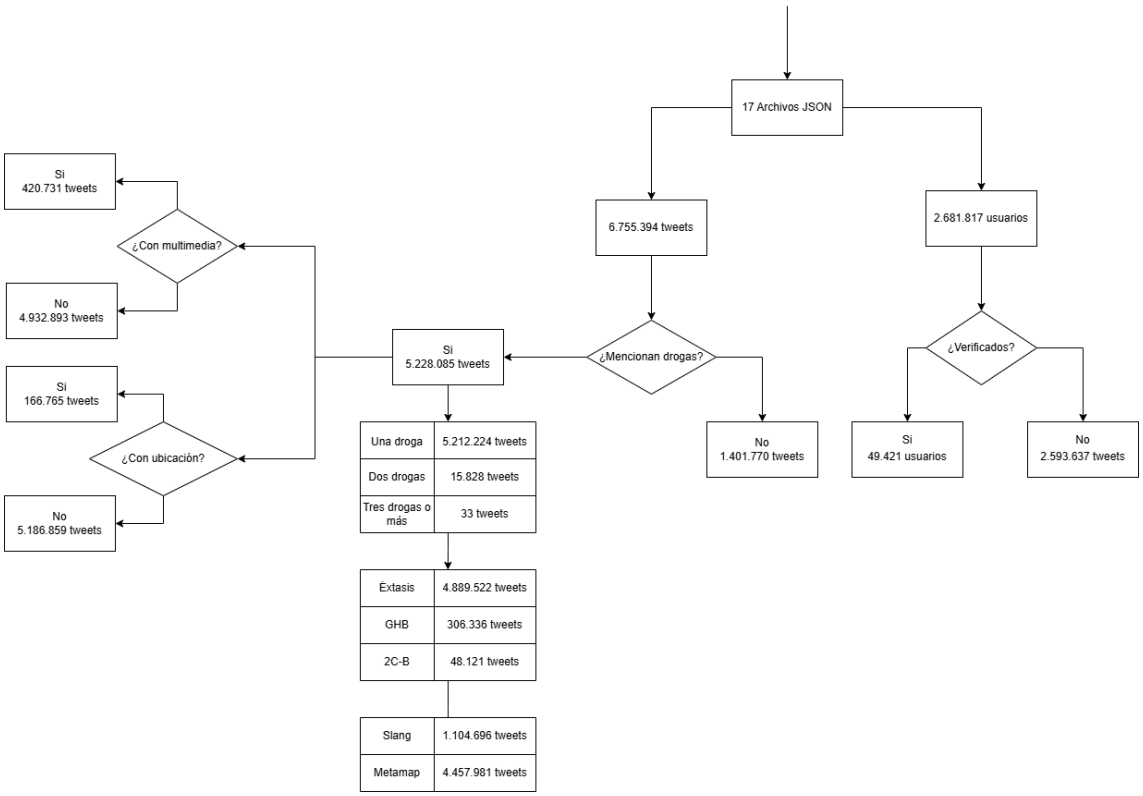


Figura 5. Diagrama de flujo de procesamiento de tweets desde archivos JSON.

6.1 Tweets

El número total de tweets que contiene el *dataset* es de 6.755.394, además podemos saber que, de esos tweets, 5.212.224 tweets mencionan una sola droga, esto sugiere que la mayoría de las conversaciones en Twitter están centradas en una sustancia específica, 15.828 tweets mencionan dos drogas, aunque es un porcentaje bajo, podrían ser de casos de comparación, consumo combinado, etc, y tan solo 33 tweets mencionan 3 o más drogas.

En la **Figura 6** se puede ver el número total de tweets por droga, 4.889.522 mencionan el éxtasis, lo que constituye un 93,24%, es decir, casi el 100% de los tweets mencionan esta droga, esto es debido a que tiene un uso recreativo más extendido o su presencia en la cultura popular es más alta que las otras drogas, el 0,92% de los tweets mencionan el 2C-B y el 5,84% de los tweets mencionan el GHB.

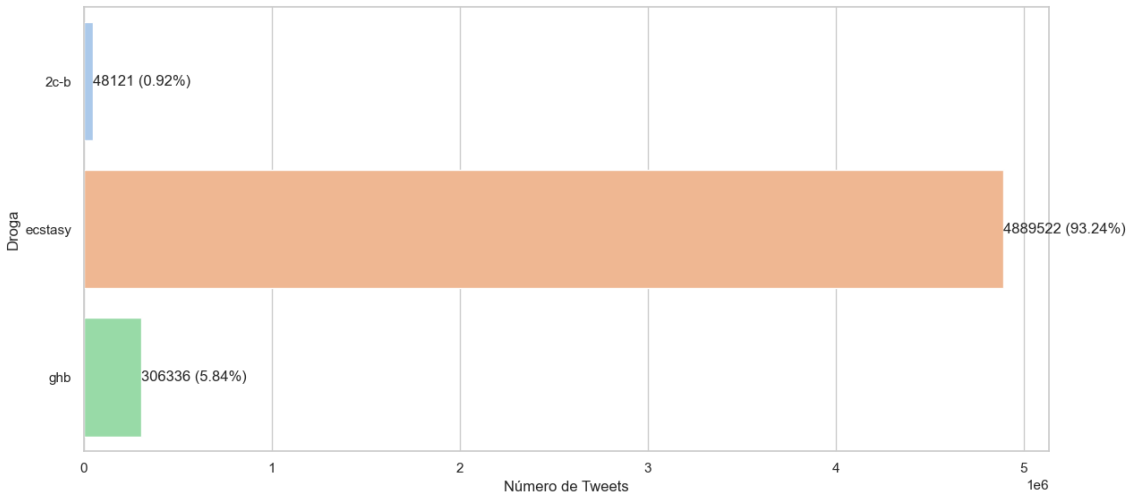


Figura 6. Distribución de tweets por droga.

El análisis de palabras clave revela patrones lingüísticos distintivos para cada droga se puede ver en el **Figura A.1 del Anexo III**. Para el GHB, términos como "date rape drug" y "easy lay" evidencian asociaciones con agresiones sexuales. El éxtasis muestra una variedad terminológica más amplia ("ecstasy", "MDMA", "XTC") con connotaciones positivas como "love doctor" o "moon rock". El 2C-B presenta términos técnicos ("nexus", "2c-b") y coloquiales ("pink cocaine", "tucibi") que reflejan su menor popularidad, pero creciente presencia intercultural. En la **Figura 7** presenta un ejemplo de tweet que forma parte del conjunto de publicaciones relacionadas con la droga éxtasis, dado que incluye el término "MDMA", identificado como una palabra clave asociada a esta sustancia.

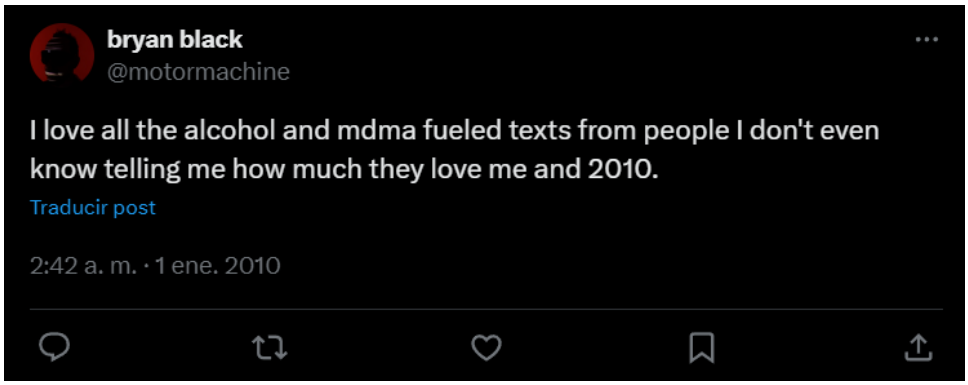


Figura 7. Tweet que menciona la palabra "MDMA" [2].

El análisis terminológico revela no solo la prevalencia del éxtasis frente a otras sustancias, sino también diferencias significativas en la percepción social: el GHB se asocia frecuentemente con contextos de vulnerabilidad y agresión, mientras que el éxtasis presenta un imaginario más positivo y recreativo. El 2C-B, aunque menos mencionado, muestra una terminología que sugiere un conocimiento más técnico o especializado entre sus usuarios.

6.2 Temporalidad

La evolución temporal, que podemos ver en la **Figura 8**, muestra un crecimiento exponencial en menciones sobre drogas entre 2007 y 2012, cuando alcanza su máximo con más de 1 millón de tweets. Este período coincide con la consolidación de Twitter como plataforma global. Posteriormente se observa un descenso constante hasta 2019, posiblemente debido a cambios en las políticas de moderación de contenido o en los patrones de uso de la plataforma, aunque en este último año podemos apreciar un pequeño crecimiento.

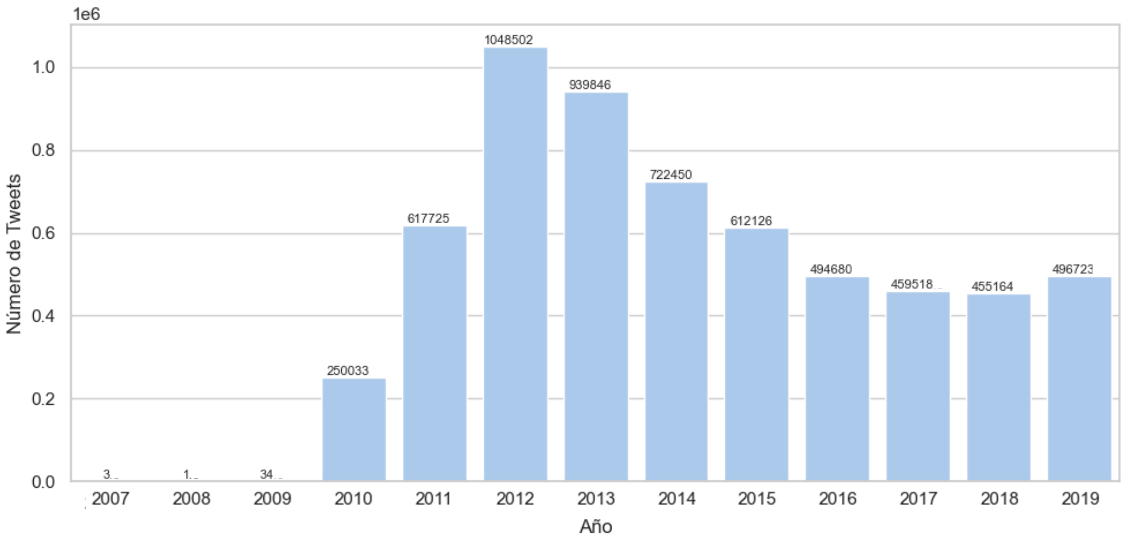


Figura 8. Distribución de tweets por año.

El análisis mensual de la **Figura 9** revela patrones estacionales con picos en junio y diciembre, que podrían correlacionarse con períodos vacaciones y

[2] https://x.com/motormachine/status/7254797006

festividades donde el consumo recreativo suele aumentar. Estos picos son especialmente pronunciados en los años 2012 y 2013, coincidiendo con el período de mayor actividad general sobre el tema.

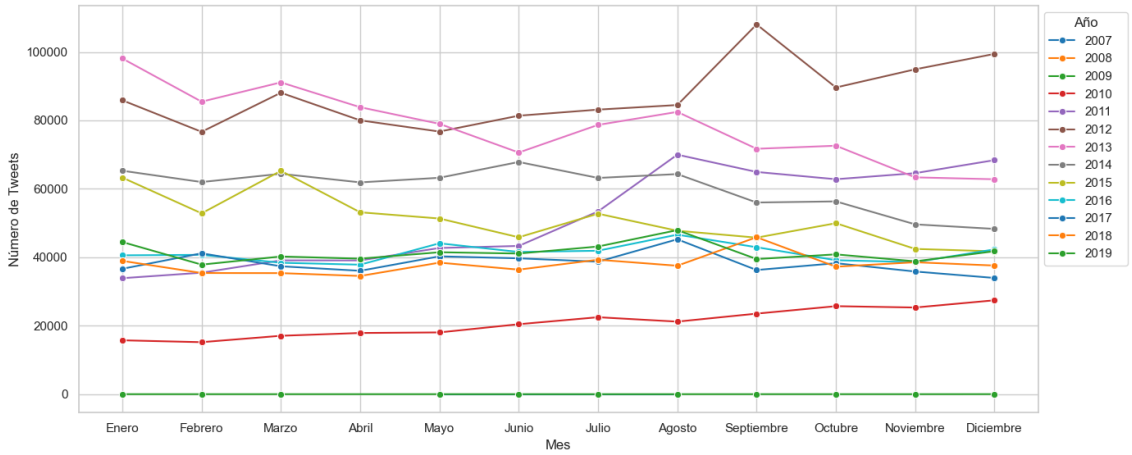


Figura 9. Distribución de tweets por mes.

El análisis anual por sustancia, mostrado en la **Figura A.2 del Anexo III**, indica que cada droga sigue patrones únicos: el 2C-B experimenta un aumento significativo en 2016 que disminuye bruscamente después, posiblemente relacionado con eventos mediáticos específicos. El éxtasis y el GHB siguen patrones más similares al comportamiento general, con máximos en 2012 y fluctuaciones posteriores menos pronunciadas.

La distribución horaria, que se puede ver en la **Figura A.4 del Anexo III**, es similar para las tres sustancias, con menor actividad durante la madrugada y un incremento progresivo que alcanza su máximo entre las 18:00 y 22:00 horas, coincidiendo con los horarios de mayor actividad general en redes sociales.

El análisis temporal revela una evolución no lineal en las conversaciones sobre drogas en Twitter, con un auge claro entre 2010-2013 seguido de una disminución gradual. Los patrones estacionales y horarios sugieren una correlación entre la actividad en redes sociales sobre drogas y los ciclos de ocio y socialización, con especial intensidad en períodos festivos y horarios vespertinos/nocturnos.

6.3 Geolocalización

La distribución geográfica que se puede ver en la **Tabla 6** muestra una clara concentración de tweets en países de habla inglesa, con Estados Unidos liderando (92.812 tweets) seguido por Reino Unido (21.535 tweets). Esta distribución refleja tanto la prevalencia del inglés en Twitter como posibles diferencias culturales en la apertura para discutir públicamente temas relacionados con drogas. Canadá ocupa un distante tercer lugar con 4.326 tweets, mientras que países como India, Filipinas y Brasil presentan cifras significativamente menores.

Tabla 6. Número de tweets por país.

País	Nº total de tweets
Estados Unidos	92.812
Reino Unido	21.535
Canadá	4.326
India	2.590
República de Filipinas	2.158
Brasil	2.097
Irlanda	2.083
Indonesia	1.884
Australia	1.789
South Africa	1.718

El mapa de calor global de la **Figura 10** refuerza esta observación, mostrando áreas de alta densidad en América del Norte, Reino Unido e India, con focos secundarios en Australia y partes de África. Las regiones con menor actividad, como Rusia y gran parte de África, pueden reflejar tanto menor adopción de la plataforma como diferencias en la regulación de contenidos relacionados con drogas.

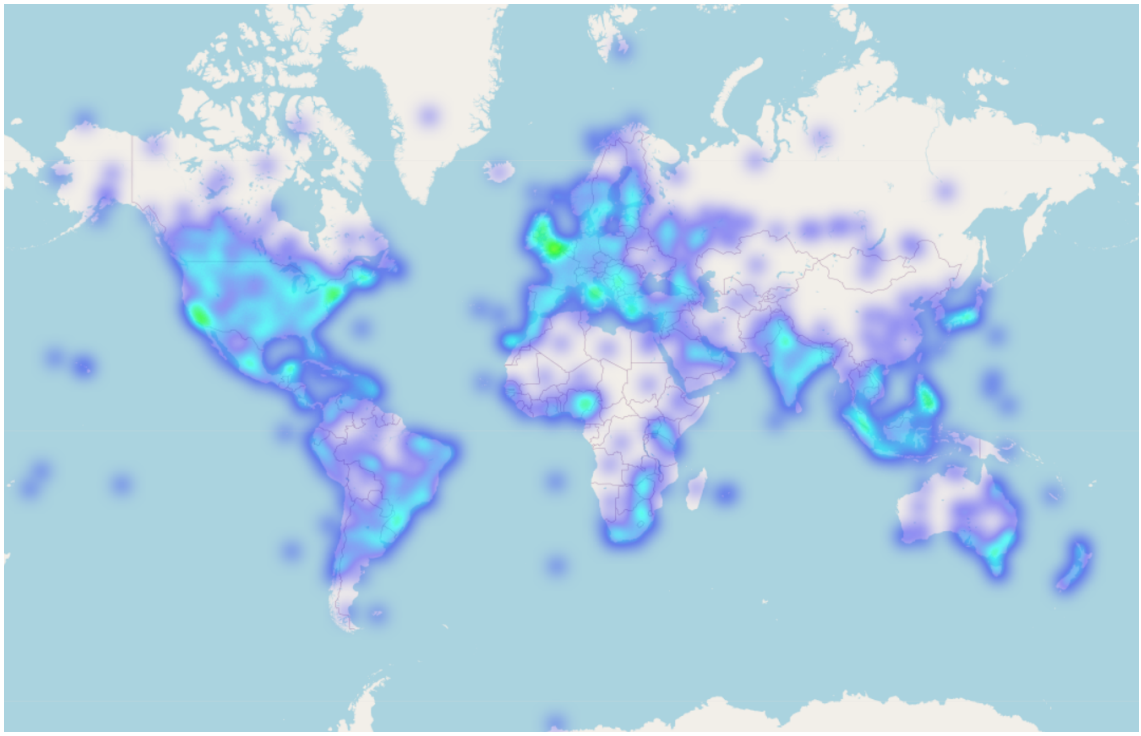


Figura 10. Mapa de calor de la distribución global de los tweets.

El análisis por sustancia, que se puede ver en la **Figura 5 del Anexo III**, confirma que el éxtasis domina en todos los países, representando aproximadamente el 92,65% de las menciones en Estados Unidos y proporciones similares en otros territorios. Esta consistencia sugiere que la

prevalencia del éxtasis en el discurso digital sobre drogas es un fenómeno global y no limitado a contextos culturales específicos.

En conclusión, la distribución geográfica revela un claro predominio de los países anglosajones en las conversaciones sobre drogas en Twitter, posiblemente reflejando factores lingüísticos, culturales y de adopción tecnológica. El dominio específico del éxtasis es consistente a nivel global, sugiriendo que esta sustancia ha alcanzado un nivel de presencial cultural transnacional superior a las otras drogas analizadas.

6.4 Usuarios

Del total de 2.604.353 tweets analizados, provenientes de diferentes autores, 2.302.839 mencionan drogas, lo que evidencia la relevancia del tema. Significativamente, de los 47.171 tweets publicados por cuentas verificadas, 31.606 (67%) mencionan drogas, indicando que estas cuentas tienen un papel proporcionalmente importante en la conversación.

El análisis de usuarios muestra una concentración notable de publicaciones en ciertas cuentas. Entre los usuarios no verificados, destacan 'Kamasutra_neue' y 'ASSENTCitizens' con 25.070 y 11.313 tweets respectivamente, y en cuanto a cuentas verificadas, 'urbandictionary' lidera con 8.618 publicaciones, seguida por 'MAPS' con 2.182 tweets. Las tablas completas con las 5 cuentas verificadas y no verificadas que publicaron el mayor número de tweets se presentan en el **Anexo III**.

En la **Figura 11**, se muestra un ejemplo de un tweet publicado por el usuario verificado, 'MAPS'. Como se puede observar, el contenido del tweet describe una teleconferencia en la que se discutieron diseños para un estudio relacionado con el uso de MDMA. Al provenir de una cuenta verificada, el mensaje adquiere un tono informativo y refleja una conversación profesional sobre el uso de esta sustancia en un contexto clínico.

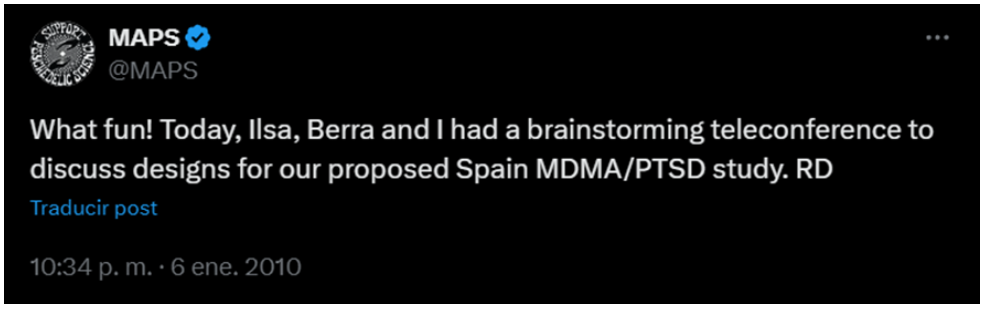


Figura 11. Tweet publicado por el usuario verificado MAPS [3].

En resumen, de un total de 2.681.817 de usuarios, un número limitado de cuentas no verificadas y verificadas muestra una actividad especialmente alta en la conversación sobre drogas en Twitter. Esta concentración sugiere que dichas cuentas tienen un papel destacado en la difusión y posible legitimación de ciertos discursos sobre drogas.

[3] https://x.com/MAPS/status/7454513970

6.5 Referencias

El análisis del tipo de interacciones la **Figura A.6 del Anexo III**, muestra un claro predominio de la categoría “replied_to” con 1.422.066 menciones, lo que indica que las conversaciones sobre drogas tienden a generar diálogo directo entre usuarios. En contraste, tanto las citas (“quoted”) como los *retweets* son significativamente menos frecuentes.

La predominancia de respuestas directas sobre otros tipos de interacción sugiere que las conversaciones sobre drogas en Twitter son más dialógicas que unidireccionales, generando comunidades de discusión en lugar de simples cadenas de difusión de información. Este fenómeno podría indicar que el tema de las drogas suscita debate y opiniones contrapuestas más que consenso y redistribución de contenido.

6.6 Engagement

El análisis de engagement mostrado en la **Figura 12** revela patrones diferenciados según la sustancia: mientras que el éxtasis concentra el mayor volumen total de interacciones (coherente con su alta frecuencia general), sorprendentemente, el 2C-B muestra los promedios más altos de ‘likes’ (7,48) y ‘retweets’ (2,75) por publicación. Esto sugiere que, a pesar de su menor presencia, el contenido sobre 2C-B genera un interés proporcionalmente mayor entre los usuarios que lo encuentran.

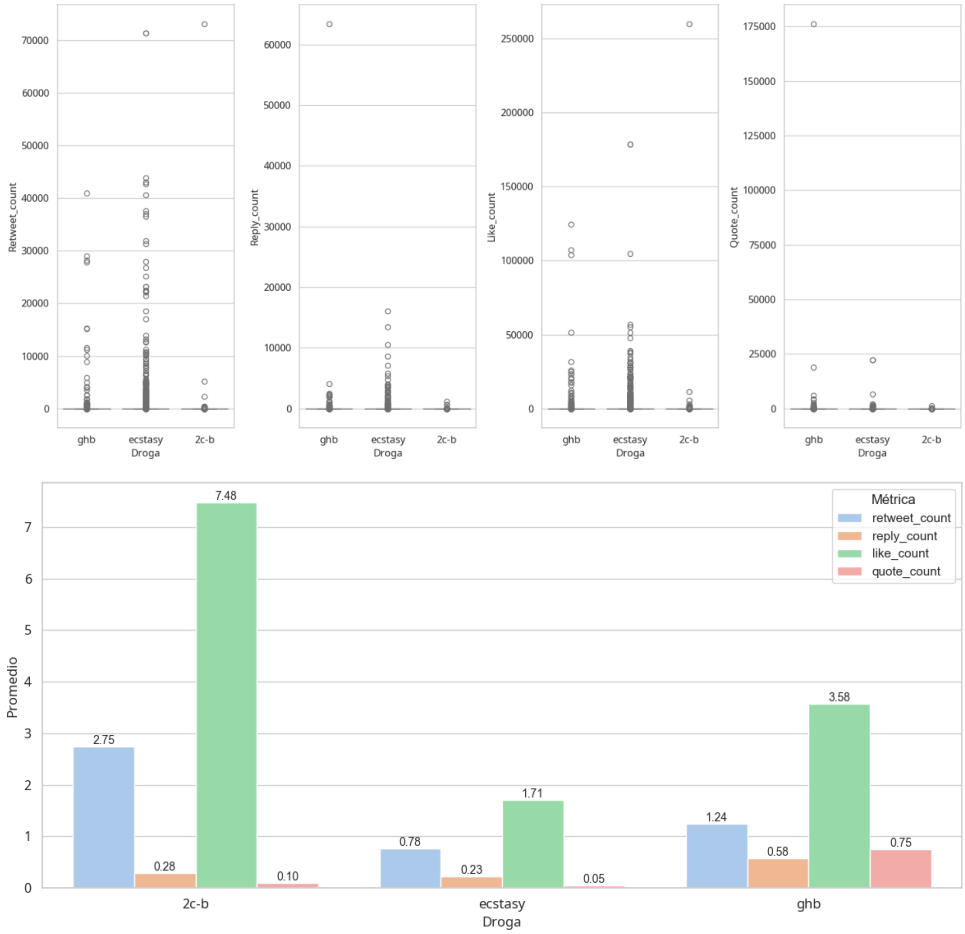


Figura 12. Distribución y promedio de interacciones por droga.

Por otra parte, la **Figura 13** muestra la matriz de correlación entre tipos de interacción. Se observa una fuerte asociación entre 'likes' y 'retweets' (0,85), así como entre 'respuestas' y 'citas' (0,90), sugiriendo que estas interacciones suelen darse conjuntamente. En contraste, la baja correlación entre 'citas' y 'retweets' (0,25) indica que estos comportamientos responden a motivaciones diferentes.

	retweet_count	reply_count	like_count	quote_count
retweet_count	1	0,43	0,85	0,25
reply_count	0,43	1	0,41	0,9
like_count	0,85	0,41	1	0,39
quote_count	0,25	0,9	0,39	1

Figura 13. Matriz de correlación de las interacciones.

Adicionalmente, otros factores también influyen en el nivel de *engagement*. Por ejemplo, las cuentas verificadas tienden a recibir significativamente más *likes* que las no verificadas, y aquellas con más de un millón de seguidores muestran el mayor promedio de interacciones. Estos patrones, desarrollados en detalle en la **Figura A.7 del Anexo III**, reflejan cómo la credibilidad percibida y el alcance de la cuenta inciden directamente en la visibilidad del contenido.

Asimismo, la presencia de contenido multimedia (como imágenes o videos) se asocia con un *engagement* promedio notablemente superior al de los tweets sin elementos visuales. Esta tendencia, presentada en la **Figura A.8 del Anexo III**, resalta la importancia de la presencia de elementos visuales para captar la atención de los usuarios y fomentar interacciones, lo que es especialmente relevante en temas de las drogas, donde el impacto visual puede ser clave para generar interés.

6.7 Slang

El análisis de los términos de jerga (*slang*) en tweets sobre el consumo recreativo de drogas explora el lenguaje informal usado en Twitter para describir sustancias, experiencias y contexto asociados. La jerga actúa como un marcador de identidad dentro de comunidades digitales, reflejando actitudes y pertenencia a subculturas específicas [15].

Del total de tweets analizados, solo un 16,4% contiene términos de jerga, frente al 83,6% restante que emplea un lenguaje más formal o directo. Este resultado sugiere que, aunque la jerga es un recurso expresivo relevante, su uso es limitado, posiblemente por razones como la necesidad de claridad, el deseo de alcanzar una audiencia más amplia o la intención de evitar malentendidos. La proporción detallada de tweets con y sin jerga se presenta en la **Figura A.10 del Anexo III**.

Cada término de *slang* fue evaluado por tres anotadores, quienes debían clasificar si su uso en un tweet sugería que el usuario estaba expresando una experiencia positiva, negativa o ambigua en relación con el consumo recreacional de drogas. Los términos se clasificaron como de connotación positiva, negativa o dependiente del contexto. Los resultados muestras que la categoría más frecuente fue la connotación dependiente del contexto (50,9%), lo que significa que el sentido de muchos términos cambia dependiendo de cómo aparece en el tweet, y no puede entenderse correctamente si se analiza de forma

aislada. Por otro lado, un 39% de los términos se consideran negativos, y solo un 10,1% positivos. Esto indica que, aunque algunos términos pueden tener diferentes significados según el contexto, en general el lenguaje usado tiende a resaltar los efectos negativos del consumo de drogas más que los positivos. La distribución detallada de estas categorías puede consultarse en la **Figura A.11 del Anexo III.**

Los términos de jerga más frecuentes y la relación entre su uso y el estado de verificación de los usuarios se incluyen en el **Anexo III**. En resumen, estos datos indican que el uso de jerga es más común entre usuarios no verificados, lo que podría asociarse con una mayor cercanía a comunidades informales o menor autocensura en la expresión.

El estudio de la jerga en tweets revela su función como marcador identitario dentro de subculturas digitales. Aunque no es dominante en cantidad, su uso está fuertemente ligado al contexto y se asocia principalmente con efectos negativos. Esto muestra cómo el lenguaje informal permite a los usuarios narrar experiencias y expresar posicionamientos sobre el consumo de drogas, proporcionando una ventana a las dinámicas socioculturales.

6.8 Metamap

Este análisis emplea MetaMap, una herramienta desarrollada por la *National Library of Medicine* (NLM) de Estados Unidos, para identificar términos biomédicos con conceptos del UMLS, estructurando así la información expresada en lenguaje coloquial presente en los tweets, dentro de categorías médicas estandarizadas, lo que facilita el análisis semántico de grandes volúmenes de texto no estructurado [16], [17].

MetaMap funciona mediante técnicas de NLP y mapeo léxico, lo que permite reconocer términos médicos incluso cuando aparecen con abreviaciones, errores ortográficos o jerga informal. Esta capacidad resulta especialmente útil en el análisis de contenido generado por usuarios en redes sociales, donde el lenguaje no sigue convenciones formales. Sin embargo, en este trabajo no se emplean directamente estas técnicas, en su lugar, se aprovechan *matched words* identificadas por MetaMap, es decir, las palabras o expresiones exactas del texto que fueron asociadas a conceptos del UMLS.

A continuación, se presentan los Tipos Semánticos (TUIs) identificados por MetaMap en los tweets analizados. Estos TUIs reflejan una amplia gama de categorías, tales como sustancias químicas, enfermedades y procedimientos médicos, lo que pone en evidencia la capacidad de MetaMap para estructurar información médica incluso en contextos informales como Twitter. Entre esta diversidad, destacan particularmente los TUIs asociados a sustancias farmacológicas (‘phsu’) y procesos mentales (‘menp’), ya que son los que aparecen en mayor número de tweets, sugiriendo que la mayoría de las publicaciones contienen referencias explícitas o implícitas a los efectos médicos y farmacológicos del consumo de drogas. La **Figura 14** ofrece una representación combinada de los TUIs detectados, mostrando tanto los colores asignados a cada categoría como su frecuencia de aparición en los tweets.

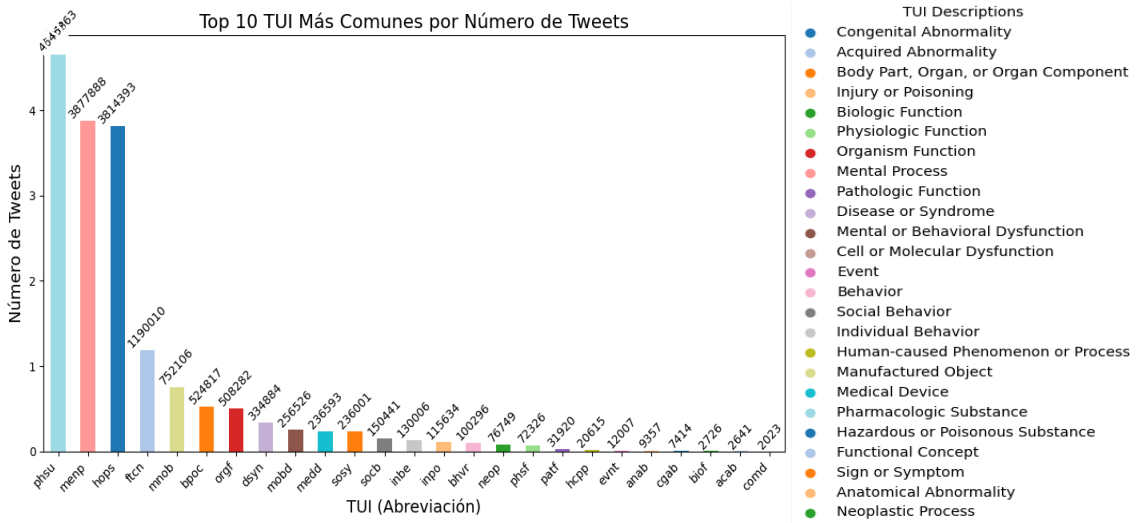


Figura 14. Distribución de TUIs: colores y frecuencia en tweets.

Además, al clasificar los TUIs según si están asociadas a aspectos psicológicos o físicos, se observa que un predomino claro de los primeros. En concreto, el 79,2% de los conceptos extraídos se vinculan con proceso mentales o emocionales, frente a un 20,8% relacionados con síntomas o condiciones físicas. Esta distribución, ilustrada en la **Figura A.14 del Anexo III**, sugiere que las conversaciones en Twitter sobre el consumo recreativo de drogas tienden a centrarse más en las consecuencias emocionales y sociales que en los efectos fisiológicos.

Por su relevancia temática, se incluye el análisis específico de los términos 'menp' más frecuentes por cada droga. En el caso del éxtasis, destacan términos vinculados a emociones positivas como "love" y "feelings", mientras que el GHB muestra una combinación entre efectos cognitivos y negativos, como "thinking" y "fear". Por su parte, los tweets sobre 2C-B tienden a incorporar referencias a experiencias sensoriales y oníricas, como "dreams". Estos resultados, reflejados en la **Figura A.15 del Anexo III**, refuerzan la idea de que el impacto psicológico ocupa un lugar central en el discurso digital sobre estas sustancias.

El uso de MetaMap ha demostrado ser eficaz para extraer conceptos biomédicos a partir de lenguaje informal, revelando patrones discursivos que giran en torno a enfermedades, efectos psicológicos y componentes químicos. La predominancia de procesos mentales destaca el enfoque emocional y conductual de las conversaciones, mientras que la detección de término técnicos valida el potencial de estas herramientas para analizar fenómenos sociales desde una perspectiva médica. Así, MetaMap permite tender un puente entre la espontaneidad del lenguaje en redes sociales y la precisión terminológica del ámbito biomédico.

7 Modelado

7.1 Preparación del dataset

La construcción de un *dataset* fiable y correctamente anotado constituye un paso clave en el desarrollo de modelos de procesamiento del lenguaje natural, tanto para su entrenamiento como para su evaluación. El modelado del *dataset* se inició a partir de dos fuentes principales: la tabla de términos *slang* y la tabla de conceptos extraídos mediante MetaMap. Ambas contenían anotaciones preliminares que requerían ser consolidadas a fin de determinar la polaridad final (positiva, negativa o contexto) de cada término o concepto.

7.1.1 Establecimiento del acuerdo de anotación

Dada la naturaleza individual de las anotaciones, es decir, una decisión por término o concepto, no resultaba adecuado utilizar métricas de concordancia global como el coeficiente *Kappa de Fleiss*, que están diseñadas para evaluar el acuerdo entre múltiples anotadores sobre un conjunto común de ítems [18]. En su lugar, se aplicó un criterio de mayoría ponderada para determinar la polaridad final de cada entrada.

Concretamente, para cada término o concepto anotado, se contabilizó la frecuencia de cada categoría de polaridad asignada por los anotadores (positivo, negativo o contexto). A partir de estas frecuencias, se calculó el porcentaje correspondiente a la categoría más votada. Si esta superaba el umbral del 60%, se consideró dicha categoría como la anotación final. En caso de que ninguna categoría alcanzara ese umbral, asignó el valor `NULL`, indicando la ausencia de un consenso suficiente.

Este procedimiento garantizó que solo se conservaran en las siguientes etapas del análisis aquellas entradas con un grado razonable de acuerdo entre los anotadores. Tras aplicar este criterio, se actualizaron las tablas correspondientes de la base de datos (‘slang’ y ‘metamap’) incorporando una nueva columna denominada ‘final_annotation’, que recoge la polaridad consolidada (positivo, negativo, contexto o `NULL`) para cada término o concepto. Esta anotación final sirvió como base para las tareas de clasificación de tweets en función del contenido léxico.

7.1.2 Clasificación de Tweets según Polaridad Léxica

Una vez consolidadas las anotaciones finales de los términos *slang* y conceptos extraídos mediante MetaMap, se procedió a la clasificación léxica de los tweets en función de la polaridad de los elementos lingüísticos que contenían. Este proceso se llevó a cabo de forma independiente para las tablas ‘tweet_slang’ y ‘tweet_metamap’, que vinculan cada tweet con los términos o conceptos relevantes previamente identificados.

Para cada tweet, se aplicó una heurística de agregación de polaridades que considera las distintas ocurrencias de términos de jerga o conceptos identificados por MetaMap dentro del mismo tweet. Primero, se agruparon todas estas apariciones en cada tweet para luego asignarles una puntuación numérica basada en su anotación consolidada:

- +1 si la polaridad era positiva
- −1 si la polaridad era negativa
- 0 si la anotación correspondía a contexto (valor neutro)

La lógica continuó con el cálculo de una puntuación total por tweet, sumando las puntuaciones asignadas a todos los términos o conceptos relacionados. Finalmente, esta puntuación global sirvió para definir la polaridad del tweet según la siguiente heurística:

- Si la puntuación total era mayor que 0, el tweet se clasificó como positivo.
- Si la puntuación total era menor que 0, se clasificó como negativo.
- Si la puntuación era igual a 0, se clasificó como contexto.

Esta clasificación se almacenó en una nueva columna añadida a las tablas 'tweet_slang' y 'tweet_metamap', permitiendo así una primera aproximación al análisis de sentimientos basada exclusivamente en la polaridad léxica de los elementos reconocidos en cada tweet.

7.1.3 Consolidación final del *dataset*

La fase final en la construcción del *dataset* consistió en la consolidación y filtrado riguroso de los tweets previamente clasificados a partir de las fuentes léxicas *slang* y MetaMap. El objetivo fue generar la tabla 'classified_tweet', que constituye el conjunto de datos definitivo utilizado para las tareas de modelado, asegurando una elevada confiabilidad en la polaridad asignada a cada tweet.

La inclusión de un tweet en esta tabla final estuvo sujeta a dos criterios restrictivos. Primero, el tweet debía estar presente y haber sido clasificado tanto en 'tweet_slang' como en 'tweet_metamap', lo que asegura que el contenido del tweet refleje simultáneamente términos de jerga y conceptos biomédicos, aportando así una mayor riqueza semántica. En segundo lugar, la clasificación de polaridad (ya sea positiva o negativa) debía coincidir en ambas tablas. Solo aquellos tweets que representaban una polaridad idéntica en ambas fuentes fueron considerados válidos para 'classified_tweet'. Aquellos con clasificaciones discordantes (por ejemplo, positiva en *slang* y negativa en MetaMap), así como los etiquetados como contexto en cualquiera de las fuentes, fueron excluidos.

Como resumen del procedimiento seguido para la construcción del *dataset* final, se presenta en la **Figura 15** un diagrama de flujo que ilustra el proceso de inclusión y exclusión aplicado a los datos provenientes de las fuentes *slang* y MetaMap. Este esquema sintetiza las distintas etapas abordadas: desde la consolidación de anotaciones léxicas individuales, la clasificación de tweets por polaridad con base en dichas anotaciones, hasta la depuración final basada en la consistencia entre ambas fuentes. Solo aquellos tweets que cumplían simultáneamente con criterio de presencia en ambas tablas y concordancia en la polaridad asignada fueron incorporados en la tabla 'classified_tweet', conformando así un conjunto de datos robusto y confiable para el modelado posterior.

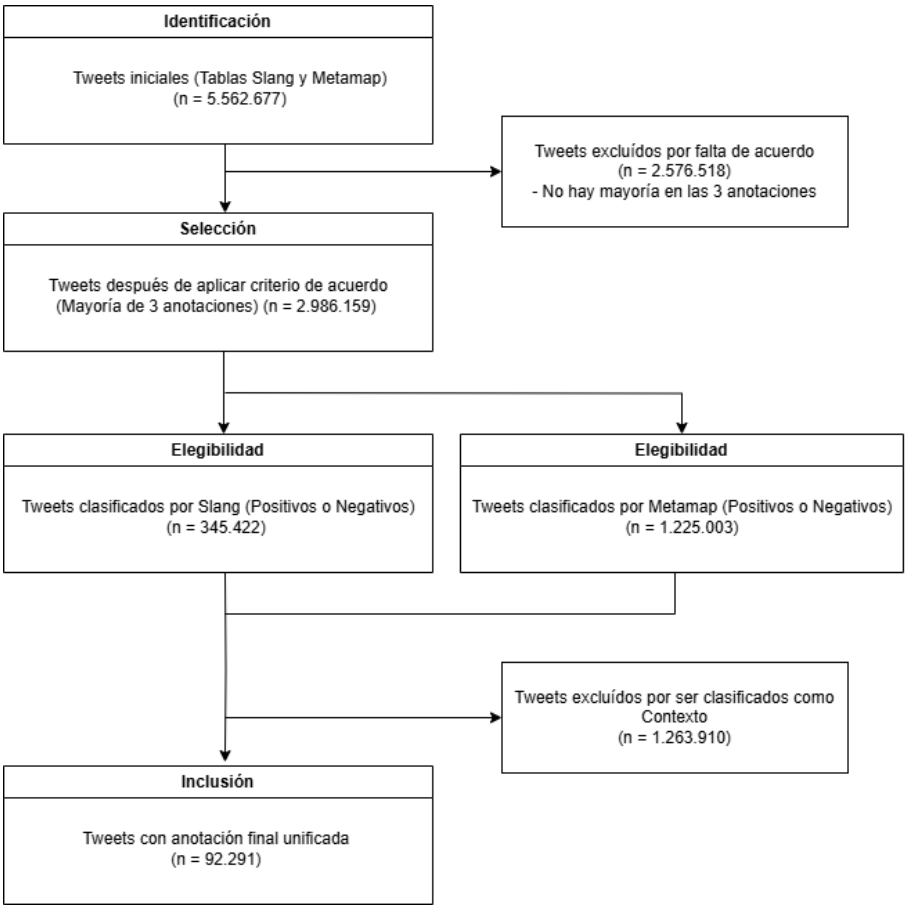


Figura 15. Diagrama de flujo del proceso de selección de tweets.

7.2 Ingeniería de características y preprocesado

7.2.1 Selección y extracción de características

Con el *dataset* 'classified_tweet' ya consolidado, se procedió a la selección y extracción de características (*feature engineering*) relevantes para el entrenamiento de modelos de clasificación. Este proceso combinó tanto metadatos del tweet y del usuario como información geográfica y la representación semántica del contenido textual. Para ellos, se utilizó una consulta SQL que se puede ver en el **Anexo IV** que integró múltiples tablas relacionadas, extrayendo un conjunto inicial de atributos que posteriormente fue exportado a un archivo Excel para su procesamiento.

La consulta SQL recuperó características agrupadas en tres grandes bloques:

Características del tweet:

- 'text': Contenido textual del tweet.
- 'media': Indicador binario de la presencia de contenido multimedia.
- 'mention' y 'reference': Indicadores de si el tweet contiene menciones a otros usuarios o si hace referencia a otro tweet (como una respuesta o *retweet*)
- Métricas de interacción: 'like_count', 'retweet_count', 'reply_count' y 'quote_count'
- Presencia de palabras clave específicas asociadas a sustancias de interés: 'mention_ghb', 'mentions_ecstasy', 'mentions_2cb'.

Características geográficas:

- Indicadores binarios que identifican la región geográfica desde donde se publicó el tweet ('is_europe', 'is_africa', 'is_asia', etc.) según el código de país ('country_code') presente en la tabla 'ubication'.
- 'user_location': Indicador de si la ubicación textual del usuario está disponible.

Características del usuario:

- 'user_verified': Indica si la cuenta está verificada.
- Métricas de actividad y visibilidad del usuario: 'user_followers', 'user_following', 'user_tweet_count', 'user_listed_count'.

Finalmente, la columna 'classification' de la tabla 'classified_tweet', con valores "positivo" o "negativo", se utilizó como variable objetivo (*target*) para los modelos de clasificación.

7.2.2 Preprocesamiento y vectorización del texto

El contenido textual de los tweets constituye una de las variables más informativas, pero también requiere un tratamiento específico antes de ser utilizado por algoritmos de aprendizaje automático. Para ello, se implementó un pipeline de preprocesamiento y vectorización mediante el modelo Word2Vec, con el objetivo de capturar el significado semántico y las relaciones contextuales entre palabras [19].

El *dataset* exportado fue cargado en un DataFrame de pandas, donde se eliminaron todas las filas con valores nulos en la columna 'text' y posteriormente aquellas con cualquier otro valor nulo. A continuación, se aplicó una función de preprocesamiento al texto que incluyó:

- Conversión a minúsculas.
- Eliminación de URLs, menciones (@), hashtags (#), números y signos de puntuación.
- Tokenización mediante *nltk.word_tokenize*.
- Eliminación de *stop words* del idioma inglés utilizando *nltk.corpus.stopwords*.

Una vez obtenidos los tokens de cada tweet, se entrenó un modelo Word2Vec con la librería *gensim* usando los siguientes parámetros:

- 'vector_size=30': cada palabra se representa como un vector de 30 dimensiones.
- 'window=5': contexto de cinco palabras antes y después del término objetivo.
- 'min_count=2': se excluyen palabras con frecuencia menor a dos.
- 'workers=4': entrenamiento en paralelo.

Para representar cada tweet como un vector único, se calculó el promedio de los vectores Word2Vec correspondientes a sus palabras. En casos donde ningún término del tweet estuviera en el vocabulario del modelo, se asignó un vector nulo de 30 dimensiones.

Los vectores generados ('w2v_0' a 'w2v_29') fueron integrados con el resto de las características numéricas obtenidas previamente. A continuación, todas las variables numéricas fueron estandarizadas utilizando *StandardScaler* de *sklearn.preprocessing*, para garantizar que contribuyeran equitativamente

durante el entrenamiento del modelo. La variable 'classification' se mantuvo sin modificar.

7.2.3 Análisis de Correlación de Características

Una vez que el *dataset* fue completamente procesado, incluyendo la vectorización del texto con Word2Vec y la estandarización de todas las características numéricas, se procedió a realizar un análisis de correlación. El objetivo de este paso fue examinar las relaciones lineales entre diferentes características numéricas del *dataset* final. Identificar características altamente correlacionadas es crucial para comprender la redundancia de la información y evaluar su impacto potencial en el rendimiento y la estabilidad de los modelos de *ML.*

Para ello, se calculó la matriz de correlación de Pearson, que cuantifica la fuerza y dirección de la relación lineal entre pares de variables. Esta matriz fue guardada en un archivo Excel (`matriz_correlacion.xlsx`) para su inspección detallada. Tras analizar los resultados, se identificaron varios pares de características con una alta correlación, definida como un coeficiente de correlación absoluto superior a 0,8. Los pares más relevantes fueron:

- 'like_count' y 'retweet_count': 0,99
- 'like_count' y 'reply_count': 0,95
- 'like_count' y 'quote_count': 0,99
- 'user_followers' y 'user_listed_count': 0,91
- 'mentions_ghb' y 'mentions_ecstasy': -0,98

Ante esta fuerte colinealidad, se procedió a eliminar características redundantes con el fin de reducir la multicolinealidad, simplificar el modelo y conservar únicamente variables representativas sin pérdida significativa de información. En particular, de las métricas de interacción del tweet ('like_count', 'retweet_count', 'reply_count', 'quote_count'), se conservó solo una, ya que todas ellas mostraban una correlación extremadamente alta entre sí. De igual modo, entre 'user_followers' y 'user_listed_count', se mantuvo únicamente una de las dos.

En cuanto a la correlación negativa entre 'mentions_ghb' y 'mentions_ecstasy' (-0,98), a pesar de su magnitud, no se eliminó ninguna de estas variables. Dado que ambas representan menciones explícitas de sustancias clave para el estudio, se consideró que aportan valor informativo específico y complementario. Por ello su inclusión es justificada por ser directamente relevantes para el objetivo de clasificación, incluso estando fuertemente relacionadas entre sí desde el punto de vista estadístico.

Este proceso de depuración permitió obtener un conjunto de datos más limpio y eficiente, con menor riesgo de colinealidad, lo que puede traducirse en mejoras a la hora de interpretar los modelos y su rendimiento general. El resultado final fue un *dataset* vectorizado, estandarizado y depurado, almacenado en el archivo `dataset_final_vectorizado.xlsx`, listo para ser empleado en la fase de entrenamiento de modelos de clasificación automática.

7.3 Entrenamiento y evaluación de modelos de clasificación

Una vez completado el preprocesamiento del conjunto de datos, incluyendo limpieza, vectorización y estandarización, se procedió al entrenamiento y

evaluación de varios algoritmos de clasificación. Dado que el conjunto de datos presentaba un claro desbalance de clases (donde la clase negativa es significativamente mayor que la positiva), manejar adecuadamente este desbalance fue un aspecto clave para construir un modelo predictivo sólido y fiable. La meta principal era asegurar que el modelo no solo tuviera una alta precisión general, sino que también fuera capaz de identificar adecuadamente la clase minoritaria (tweets con polaridad positiva).

El *dataset* procesado se dividió en dos subconjuntos, entrenamiento con un 80% de los datos, equivalente a 73.833 tweets, y prueba con un 20%, de forma estratificada (`stratify=y`) para asegurar que la proporción de clases se mantuviera consistente en ambos subconjuntos. Esto es crucial en *datasets* desbalanceados para garantizar que tanto el entrenamiento como la evaluación sean representativos.

Para cada modelo y estrategia de balanceo, se realizó una búsqueda de hiperparámetros utilizando *RandomizedSearchCV*. Esta técnica explora un número fijo de combinaciones de hiperparámetros, seleccionadas aleatoriamente dentro de las distribuciones definidas. La métrica de optimización principal fue el *F1-Score*, seleccionada por su competencia en problemas de clasificación con desbalance de clases, ya que combina precisión y exhaustividad en una sola métrica dando una medida equilibrada del rendimiento del modelo [20].

Con el objetivo de obtener una evaluación más precisa del rendimiento de los modelos y reducir el riesgo de sobreajuste durante la búsqueda de hiperparámetros, se utilizó una estrategia de **validación cruzada anidada y estratificada** (*StratifiedKFold*) con 5 pliegues (`n_splits=5`), organizada en dos niveles. El bucle interno se usó para encontrar los mejores hiperparámetros mediante *RandomizedSearchCV*, mientras que el bucle externo permitió evaluar el rendimiento del modelo de forma más realista y generalizable [20],[21].

El uso de esta técnica ayuda a separar el proceso de ajuste del modelo del proceso de evaluación, evitando así resultados demasiado optimistas que pueden aparecer cuando ambos se realizar con los mismos datos. Además, al utilizar pliegues estratificados, se asegura que cada subconjunto tenga una proporción similar de clases (positivas y negativas), lo que es especialmente importante en *dataset* con clases desbalanceadas, como en este caso. En conjunto, esta metodología ofrece resultados más fiables y representativos para aplicaciones reales.

Para una evaluación exhaustiva, se consideraron diversas métricas de rendimiento, tanto para la validación cruzada (reportando el promedio y la desviación estándar) como para el rendimiento final en el conjunto de prueba:

- **Matriz de Confusión (*True negative* TN, *False Positive* FP, *False Negative* FN, *True Positive* TP)**: Proporciona un desglose detallado de los aciertos y errores de clasificación. A partir de esta matriz podemos obtener el resto de las métricas de evaluación que se presentan a continuación.
- ***Recall* (Exhaustividad)**: Proporción de la clase positiva real que fue correctamente identificada. Importante para minimizar falsos negativos.

$$\frac{TP}{TP + FN}$$

- ***Precision* (Precisión)**: Proporción de las predicciones positivas que fueron realmente correctas. Importante para minimizar falsos positivos.

$$\frac{TP}{TP + FP}$$

- ***Accuracy* (Exactitud)**: Proporción de predicciones correctas sobre el total de predicciones. Aunque útil, puede ser engañosa en datasets desbalanceados.

$$\frac{TP + TN}{TP + TN + FP + FN}$$

- **F1-Score**: Fundamental en datasets desbalanceados, ya que es una medida equilibrada de precisión y exhaustividad.

$$2 \times \frac{Precision \times Recall}{Precision + Recall}$$

- ***Area Under the Receiver Operating Characteristic curve* (AUROC)**: Mide la capacidad del modelo para distinguir entre clases. Es robusta al desbalance de clases.
- ***Area Under the Precision-Recall curve* (AUPRC)**: Especialmente informativa en escenarios de desbalance severo, ya que se enfoca en el rendimiento en la clase minoritaria.

Se evaluaron diferentes algoritmos de clasificación para comparar su desempeño en la detección de experiencias positivas y negativas. Se incluyeron modelos comunes por su efectividad, desde árboles de decisión hasta métodos más complejos como redes neuronales.

- Los árboles de decisión, más conocidos por su nombre inglés ***Decision Tree* (DT)**, son modelos no paramétricos usados en clasificación y regresión. Funcionan dividiendo progresivamente el conjunto de datos en subconjuntos más pequeños mediante reglas del tipo "si-entonces", lo que da lugar a una estructura similar a un árbol [22].
- ***Random Forest* (RF)** es un método de aprendizaje de conjunto (*ensemble learning*) que combina varios árboles de decisión, introduciendo aleatoriedad en los datos y características usadas. Esto mejora la precisión y reduce el sobreajuste en tareas de clasificación y regresión [23].
- ***Logistic Regression* (LR)** es un modelo lineal usado principalmente para clasificación binaria. Estima la probabilidad de pertenecer a una clase aplicando una función logística (*sigmoide*) a una combinación lineal de las variables de entrada [24].
- ***Extreme gradient boosting* (XGBoost)** es la versión optimizada del algoritmo de gradient boosting, conocida por su alta eficiencia y precisión. Construye árboles de forma secuencial corrigiendo errores previos, e incorpora técnicas de regularización para evitar el sobreajuste y tratar valores perdidos de forma efectiva [25].
- ***Adaptive Boosting* (AdaBoost)** es un algoritmo de *boosting* que entrena secuencialmente clasificadores simples, aumentando el peso de los errores para mejorar el siguiente modelo. Combina sus predicciones mediante un promedio ponderado [26].
- ***Bagging Classifier*** es una técnica de aprendizaje de conjunto que entrena varias versiones del mismo clasificador sobre subconjuntos aleatorios de datos para reducir la varianza y evitar el sobreajuste. Combina sus predicciones mediante votación o promedio [27].
- Las redes neuronales multicapa (***Multi-layer Perceptron* (MLP)**, por sus siglas en inglés) son un tipo de red *feedforward* compuesta por una capa

de entrada, una o más ocultas y una de salida, donde la información fluye en una sola dirección. Son capaces de aprender relaciones no lineales complejas gracias a sus funciones de activación y conexiones ponderadas entre neuronas [28].

El listado completo de modelos implementados y los espacios de búsqueda de hiperparámetros empleados durante el entrenamiento se incluyen en el **Anexo V**.

7.3.1 Estrategias de manejo del desbalance de clases

Se exploraron cuatro estrategias de entrenamiento distintas para gestionar el desbalance de clases, comparando su impacto en el rendimiento de los modelos. Los resultados detallados de las métricas obtenidas mediante la validación cruzada se presentan en el **Anexo VI**.

7.3.1.1 Entrenamiento sin técnicas de balanceo

Como punto de partida, se evaluaron los modelos entrenados sin aplicar ninguna técnica de sobremuestreo ni ajuste explícito de pesos. Esta estrategia permite establecer una línea base dese presentan los resultados de los modelos entrenados sin aplicar técnicas de sobremuestreo ni ajuste explícito de pesos. Este enfoque permite establecer una línea base de rendimiento en el *dataset* desbalanceado. La **Tabla 7** muestra los resultados obtenidos durante la evaluación final en el conjunto de prueba.

Tabla 7. Rendimiento de modelos sin oversampling (conjunto de prueba).

Método	F1	REC	PREC	ACC	TN	FN	FP	TP	AUROC	AUPRC
DT	0,7408	0,7471	0,7346	0,9732	12869	178	190	526	0,8663	0,5618
RF	0,8429	0,7357	**0,9866**	0,9859	13052	186	7	518	0,9874	0,9186
AdaBoost	0,7869	0,7343	0,8475	0,9796	12966	187	93	517	0,9768	0,8479
Bagging	0,8373	0,7386	0,9665	0,9853	13041	184	18	520	0,9720	0,8937
LR	0,8157	0,7485	0,8962	0,9827	12998	177	61	527	0,9816	0,8825
XGBoost	**0,8807**	**0,8181**	0,9536	**0,9886**	**13031**	**128**	**28**	**576**	**0,9895**	**0,9309**
MLP	0,8535	0,8196	0,8904	0,9856	12988	127	71	577	0,9889	0,9219

En este escenario, el modelo *XGBoost* presentó el mejor desempeño general, con un *F1-Score* de 0,887 ± 0,012 en validación cruzada y 0,880 en el conjunto de prueba. El modelo MLP también mostró un rendimiento muy sólido, alcanzando *F1-Score* de 0,874 en validación cruzada y 0,854 en prueba. Es importante destaca que, aun sin aplicar técnicas específicas de balanceo, estos modelos lograron métricas competitivas, comparables, y en algunos modelos superiores, a las obtenidas con enfoques como el *cost-sensitive* o el uso de SMOTE dentro de la validación cruzada. Esto sugiere que algoritmos como *XGBoost* y MLP, poseen propias para manejar cierto grado de desbalance en los datos. Asimismo, los altos valores de precisión, AUROC y AUPRC alcanzados por ambos modelos refuerzan su capacidad para identificar correctamente la clase minoritaria, incluso sin una intervención explícita para equilibrar las clases. La **Figura 16** presenta una comparación gráfica del rendimiento en este escenario.

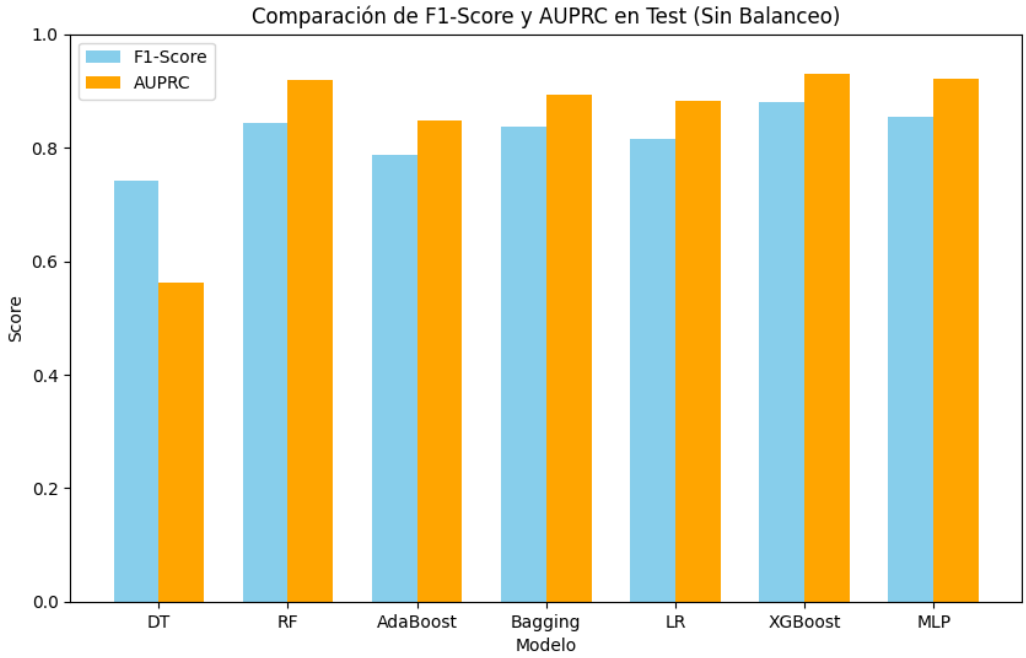


Figura 16. Comparación de los modelos en test sin técnicas de balanceo (F1-Score y AUPRC).

7.3.1.2 Modelos cost-sensitive

Esta estrategia implica la aplicación de un factor de ponderación ('class_weight' en *scikit-learn* o 'scale_pos_weight' en *XGBoost*) directamente en los algoritmos de clasificación que soportan esta funcionalidad. El objetivo es asignar un mayor "costo" o penalización a los errores de clasificación de la clase minoritaria durante el proceso de entrenamiento. Esto fuerza al modelo a prestar más atención a las muestras de la clase menos representada, reduciendo la tendencia hacia la clase mayoritaria.

Los resultados detallados de la evaluación en el conjunto de prueba se presentan en la **Tabla 8**.

Tabla 8. Rendimiento de modelos Cost-Sensitive (conjunto de prueba).

Método	F1	REC	PREC	ACC	TN	FN	FP	TP	AUROC	AUPRC
DT	0,7445	0,7286	0,7611	0,9744	12898	191	161	513	0,8581	0,5685
RF	0,8391	0,7301	**0,9865**	0,9856	13052	190	7	514	0,9872	0,9154
AdaBoost	0,7869	0,7343	0,8475	0,9796	12966	187	93	517	0,9768	0,8479
Bagging	0,8429	0,7471	0,9669	0,9857	13041	178	18	526	0,9687	0,8906
LR	0,5750	**0,9417**	0,4138	0,9287	12120	41	939	663	0,9829	0,8525
XGBoost	**0,8852**	0,8437	0,9310	**0,9888**	**13015**	**110**	**44**	**594**	0,9889	**0,9338**
MLP	0,8724	0,8451	0,9015	0,9873	12994	109	65	595	**0,9894**	0,9310

En esta estrategia, el modelo *XGBoost* volvió a destacarse como el de mejor rendimiento global, alcanzando un *F1*-Score de 0,894 ± 0,013 en validación cruzada y 0,885 en el conjunto de prueba. Este modelo demostró una alta capacidad para identificar correctamente la clase minoritaria (*recall* elevado), sin perder precisión lo que permitió mantener un buen equilibrio entre ambas métricas. Sus valores de precisión (0,931) tanto en validación como en prueba, junto con los altos AUROC (0,989) y AUPRC (0,934), refuerzan su solidez frente al desbalance de clases. Por su parte, el modelo MLP también presentó un rendimiento competitivo, con resultados cercanos a *XGBoost* tanto en F1-Score

como en AUPRC. En contraste, la Regresión Logística, a pesar de lograr un buen *recall*, mostró una precisión considerablemente más baja, lo que afectó negativamente su *F1-Score*. Esto indica que, aunque identificó muchas instancias positivas, también generó un mayor número de falsos positivos, disminuyendo su efectividad global.

Como se muestra en la **Figura 17**, el gráfico de barras permite visualizar claramente estas diferencias de rendimiento entre los modelos en términos de *F1-Score*.

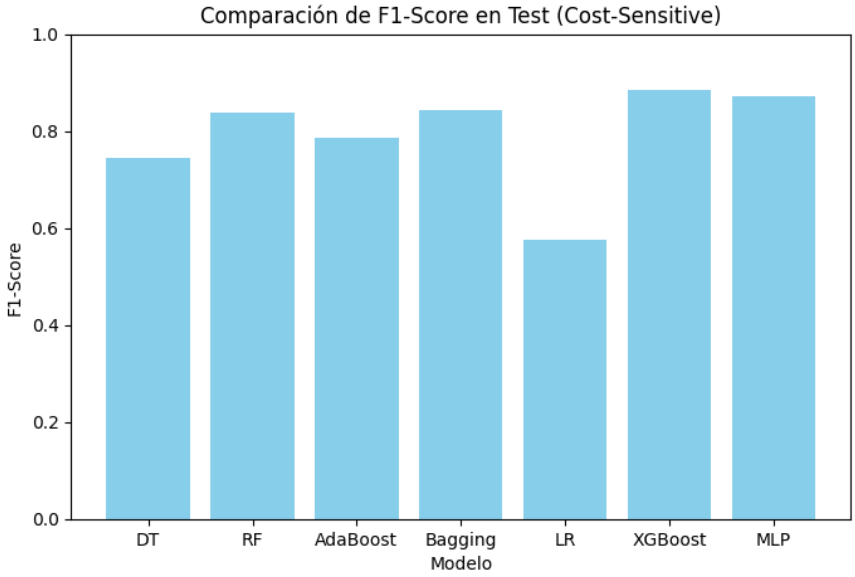


Figura 17. Comparación de F1-Score en Cost-Sensitive.

7.3.1.3 Sobremuestreo con SMOTE (aplicado previo a la validación cruzada)

Esta aproximación consistió en aplicar la técnica *Synthetic Minority Over-sampling Technique* (SMOTE) a todo el conjunto de entrenamiento antes de realizar la validación cruzada o la búsqueda de hiperparámetros. SMOTE genera nuevas muestras sintéticas de la clase minoritaria a partir de instancias reales, buscando mejorar el equilibrio entre clases. Sin embargo, esta metodología no es recomendable en la práctica, ya que introduce un riesgo elevado de fuga de información (*data leakage*) al crear datos sintéticos antes de dividir en pliegues, es posible que el modelo tenga acceso indirecto a información del conjunto de validación, lo que produce una estimación poco realista del rendimiento.

Los resultados detallados de la evaluación en el conjunto de prueba se presentan en la **Tabla 9**.

Tabla 9. Rendimiento de modelos con SMOTE Pre-CV (conjunto de prueba).

Método	F1	REC	PREC	ACC	TN	FN	FP	TP	AUROC	AUPRC
DT	0,6854	0,8125	0,5927	0,9618	12666	132	393	572	0,8912	0,4911
RF	0,8498	0,8480	0,8516	0,9846	12955	107	104	597	**0,9886**	0,9205
AdaBoost	0,5979	0,8778	0,4534	0,9396	12314	86	745	618	0,9740	0,8253
Bagging	0,8280	0,8480	0,8089	0,9819	12918	107	141	597	0,9856	0,9065
LR	0,6105	**0,9218**	0,4563	0,9398	12286	55	773	649	0,9824	0,8534
XGBoost	**0,8720**	0,8522	**0,8928**	**0,9872**	**12987**	**104**	**72**	**600**	0,9880	**0,9307**
MLP	0,8438	0,8480	0,8396	0,9839	12945	107	114	597	0,9872	0,9233

Este efecto se evidencia al comparar los resultados de la validación cruzada con los del conjunto de pruebas. El modelo *Bagging* obtuvo un *F1-Score* de 0,828 en el conjunto de prueba, una caída notable frente al 0,996 alcanzado en la validación cruzada, lo cual es una clara señal de sobreajuste (*overfitting*). A pesar de esto, algunos modelos como *XGBoost* y MLP mantuvieron un buen desempeño, con *F1-Score* de 0,8720 y 0,8438 respectivamente, además de altas puntuaciones en AUROC y AUPRC, lo que indica una cierta capacidad para generalizar. Esta diferencia de rendimiento se visualiza en la **Figura 18**, que compara los *F1-Score* obtenidos en la validación cruzada y prueba para cada modelo bajo esta estrategia.

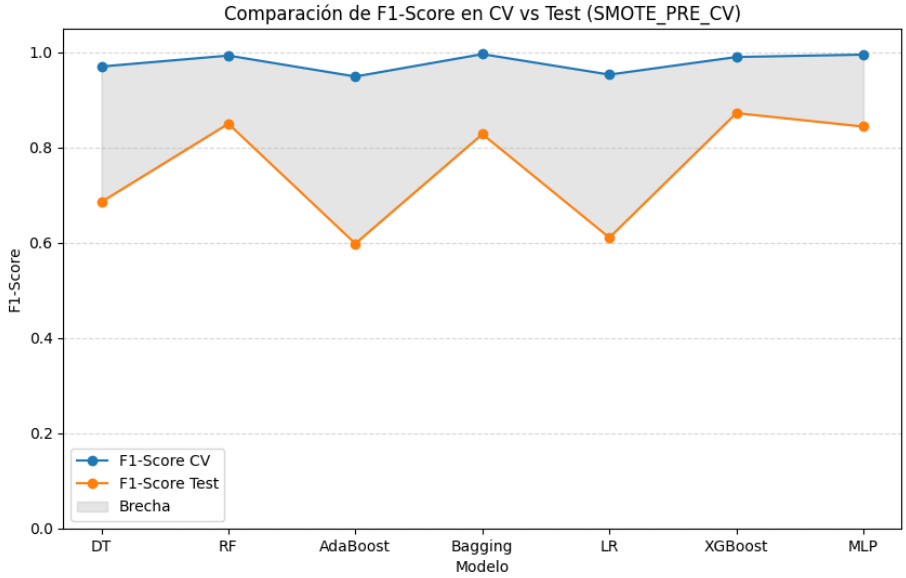


Figura 18. Comparación del F1-Score con SMOTE en CV y test

7.3.1.4 Sobremuestreo con SMOTE (Aplicado Dentro de la Validación Cruzada)

Esta estrategia representa la implementación correcta y robusta de SMOTE, asegurando que el sobremuestreo solo se realice en el conjunto de entrenamiento de cada pliegue de la validación cruzada, evitando la fuga de información. Los resultados de la evaluación en el conjunto de prueba se muestran en **Tabla 10**.

Tabla 10. Rendimiento de modelos con SMOTE In-CV (conjunto de prueba).

Método	F1	REC	PREC	ACC	TN	FN	FP	TP	AUROC	AUPRC
DT	0,6809	0,8153	0,5845	0,9609	12651	130	408	574	0,8920	0,4860
RF	0,8532	0,8508	0,8557	0,9850	12958	105	101	599	0,9885	0,9193
AdaBoost	0,6355	0,8792	0,4975	0,9484	12434	85	625	619	0,9767	0,8430
Bagging	0,8150	0,8451	0,7870	0,9803	12898	109	161	595	0,9833	0,8980
LR	0,6091	**0,9218**	0,4548	0,9394	12281	55	778	649	0,9824	0,8536
XGBoost	**0,8668**	0,8551	**0,8788**	**0,9865**	**12976**	**102**	**83**	**602**	**0,9893**	**0,9325**
MLP	0,8475	0,8451	0,8500	0,9844	12954	109	105	595	0,9876	0,9183

En esta estrategia, que proporciona las estimaciones de rendimiento más fiables, el modelo *Bagging Classifier* obtuvo el *F1-Score* más alto en validación cruzada con 0,886 ± 0,015. En el conjunto de prueba, *XGBoost* y *Bagging* se mantuvieron como los de mejor desempeño, con *XGBoost* ligeramente por delante en *F1-Score* (0,867). Ambos modelos mostraron un equilibrio sólido

entre *recall* y *precision*. Los valores de AUROC y AUPRC para *XGBoost* (0,989 y 0,933 respectivamente) demuestran su excelente capacidad discriminativa y su buen rendimiento en la clase minoritaria. Esta estrategia confirmó que SMOTE aplicado correctamente puede mejorar el rendimiento, especialmente para modelos que no tienen un manejo de pesos de clase propio (**Figura 19**).

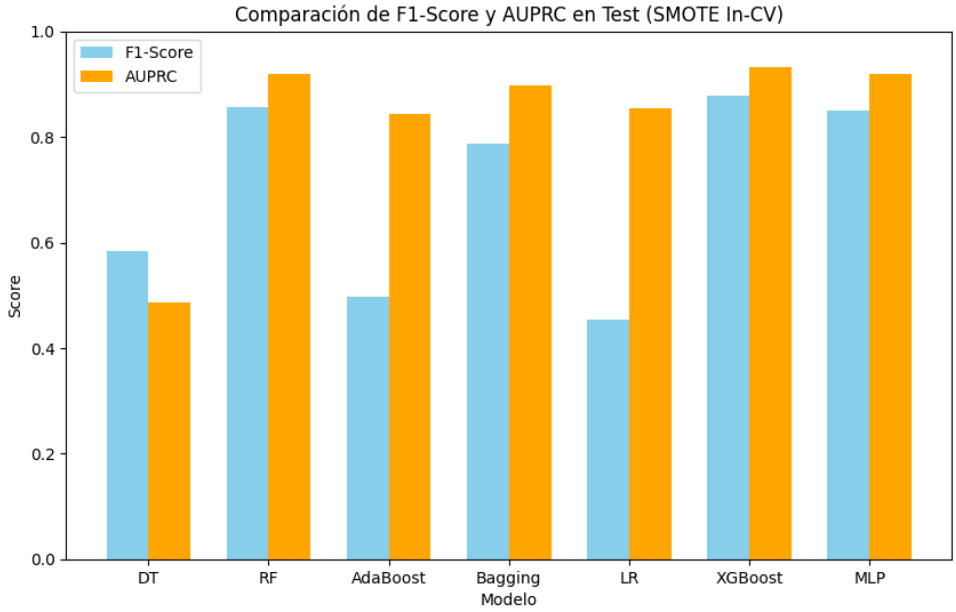


Figura 19. Desempeño de los modelos en test con SMOTE In-CV: F1-Score y AUPRC.

7.3.2 Análisis de Resultados y Comparación de Modelos

El análisis de los resultados obtenidos a partir de las cuatro estrategias implementadas para abordar el desbalance de clases permite identificar varias conclusiones relevantes y comparativas.

La aplicación de SMOTE previa al proceso de validación cruzada genera resultados que evidencian un riesgo elevado de sobreajuste. Las métricas obtenidas durante la validación cruzada, particularmente los valores de *F1-Score* cercanos a 1,0, no se replican en el conjunto de prueba, donde se observan desempeños considerablemente inferiores. Esta discrepancia resalta la importancia de aplicar técnicas de muestreo dentro del ciclo de validación cruzada para obtener estimaciones realistas del rendimiento del modelo.

La estrategia que incorpora SMOTE dentro del ciclo de validación cruzada se considera metodológicamente más rigurosa. Aunque los resultados no son tan elevados como en el caso anterior, presentan mayor consistencia entre la validación cruzada y el conjunto de prueba, reflejando de forma más precisa la capacidad del modelo para generalizar a datos no vistos.

Los modelos *XGBoost* y MLP destacan por su rendimiento robusto a lo largo de las distintas estrategias evaluadas. Ambos mantienen valores elevados y estables en métricas como *F1-Score, Recall, Precision,* AUROC, AUPRC, lo que demuestra su adecuación para tareas de clasificación con clases desbalanceadas.

Los modelos configurados con ajuste de pesos internos (*cost-sensitive*) o sin aplicar técnicas explícitas de balanceo también mostraron desempeños competitivos. En algunos casos, estos enfoques lograron resultados comparables o incluso superiores a los obtenidos mediante SMOTE dentro de

la validación cruzada. Esto sugiere que los mecanismos internos de ponderación de clases presentes en algoritmos como *XGBoost* y MLP pueden ser efectivos para reducir los efectos de desbalance.

En contextos de desbalance de clases, la elección adecuada de métricas es crucial. La exactitud (*Accuracy*) puede resultar engañosa, ya que un modelo puede alcanzar altos niveles de precisión simplemente clasificando la mayoría de las instancias como pertenecientes a la clase mayoritaria. Métricas como *Recall, F1-Score*, AUROC y AUPRC, especialmente centradas en la clase minoritaria, ofrecen una evaluación más completa y fiable del rendimiento del modelo.

XGBoost y MLP se posicionan como los modelos más prometedores para la tarea de clasificación de tweets con desbalance de clases. En cuanto a estrategias de tratamiento del desbalance, se recomienda utilizar SMOTE dentro del ciclo de validación cruzada o, alternativamente, modelos *cost-sensitive*, siempre que el algoritmo lo permita y se observe un desempeño adecuado.

8 Discusión y conclusiones

8.1 Conclusiones

Los resultados del análisis revelan que el discurso en Twitter sobre el consumo de drogas refleja percepciones culturales y sociales claramente diferenciadas según la sustancia. El éxtasis, que domina el 93,24% de los tweets analizados, se asocia principalmente con connotaciones positivas como “love doctor”, mientras que el GHB, con un 5,84%, aparece vinculado a términos negativos como “date rape drug”. Esta diferencia de imagen explica en parte la predominancia del éxtasis en las conversaciones y el mayor compromiso observado con sustancias menos frecuentes como el 2C-B, que muestra una actividad menor, pero con alta interacción, sugiriendo la presencia de comunidades de nicho en plataformas angloparlantes, especialmente en Estados Unidos y Reino Unido, donde se concentra la mayor parte de actividad. Además, la evolución temporal indica picos de publicación durante períodos vacacionales, reflejando posibles patrones sociales vinculados al consumo.

La exploración de los términos mediante MetaMap mostró una predominancia de conceptos relacionados con efectos psicológicos frente a fisiológicos, indicando que los usuarios priorizan experiencias emocionales y sociales sobre efectos físicos, un resultado relevante para diseñar estrategias de prevención y salud pública más enfocadas en el impacto psicosocial del consumo. Sin embargo, solo el 16,4% de los tweets usaron jerga, lo que sugiere una comunicación más formal o cautelosa en la plataforma.

Desde la perspectiva del modelado, algoritmos como XGBoost y MLP demostraron un desempeño robusto, alcanzando F1-Scores de hasta 0,885, especialmente cuando se aplicaron técnicas adecuadas para manejar el desbalance de clases, como modelos sensibles al costo y el uso correcto de SMOTE dentro de la validación cruzada. La detección de sobreajuste al aplicar SMOTE antes de la validación, subraya la importancia de un enfoque metodológico riguroso para obtener evaluaciones fiables. En conjunto, estos resultados no solo evidencian patrones y dinámicas en el discurso sobre drogas en redes sociales, sino que también brindan una base sólida para el desarrollo de intervenciones y análisis futuros que consideren tanto la diversidad cultural como la complejidad del lenguaje utilizado.

8.2 Limitaciones

Este estudio, aunque robusto y con un enfoque integral, enfrenta diversas limitaciones derivadas principalmente de la calidad y naturaleza de los datos, así como del alcance metodológico. Los archivos JSON obtenidos de Twitter presentaron inconsistencias y datos faltantes en campos esenciales como la ubicación geográfica o el contenido multimedia. Además, algunos tweets contenían estructuras JSON variables o atípicas, lo que obligó a implementar controles y excepciones específicos en el código de extracción e inserción para manejar estas diferencias y asegurar la integridad de la base de datos.

La dependencia exclusiva de Twitter como fuente limita la generalización de los resultados a otras redes sociales o contextos fuera de línea, dada la demografía específica, estilo lingüístico y políticas de moderación de la plataforma. El uso de una lista fija de palabras clave y jerga anotada pudo omitir discusiones con lenguaje indirecto o contenido no textual relevante, afectando la exhaustividad

de la captura de datos. El estudio se enfocó en solo tres drogas (GHB, éxtasis y 2C-B), por lo que sus conclusiones no necesariamente aplican a otras sustancias.

En cuanto al modelado, la técnica Word2vec, aunque útil, no capta matices contextuales complejos tan bien como los modelos de lenguaje más avanzado basados en transformadores (como BERT o RoBERTa). La anotación humana pese al consenso mayoritario aplicado, implica cierta subjetividad que puede influir en la fiabilidad de las etiquetas. Por último, a pesar de implementar estrategias para manejar el desbalance de clases, este persistió en el conjunto de datos, y la decisión de priorizar la calidad y coherencia de las etiquetas redujo el tamaño del conjunto disponible para el entrenamiento, representando la clásica compensación entre volumen y fiabilidad en ciencia de datos.

8.3 Líneas de trabajo futuras

Basándose en las limitaciones y observaciones de este estudio, se plantean varias direcciones prometedoras para futuros trabajos que podrían mejorar la precisión, robustez y aplicabilidad del análisis.

- **Uso de *Large Language Models* (LLMs):** El estudio actual utilizó Word2Vec para la vectorización y clasificadores tradicionales, pero se recomienda explorar LLMs como BERT o RoBERTa, que capturan mejor el contexto y los matices lingüísticos, esenciales para el análisis del lenguaje complejo y ambiguo de redes sociales [29]. Estos modelos generan representaciones dinámicas que ajustan el significado según el contexto, mejorando la clasificación precisa de textos relacionados con drogas.
- ***Fine-tuning* de LLMs:** El ajuste fino (*fine-tuning*) de LLMs preentrenados con el conjunto de datos específico permitiría especializarlos en reconocer patrones y sentimiento propios del dominio, superando a modelos genéricos [30]. El *fine-tuning* basado en *prompts*, que transforma tareas en ejercicios de "rellenar huecos", ofrece una vía eficiente para mejorar el desempeño en clasificación.
- **Clasificación *zero-shot*:** Otra línea es la clasificación *zero-shot*, donde un LLM clasifica textos sin entrenamiento específico, basándose solo en instrucciones. Esto permite evaluar rápidamente nuevos datos o términos sin etiquetado previo, aunque suele ser menos preciso que el *fine-tuning*. Modelos como GPT-4 o Claude 3 son prometedores en este enfoque, pero requieren una "ingeniería de *prompts*" cuidados para optimizar resultados [31].
- **Técnicas avanzadas de *oversampling*:** Además de SMOTE y ADASYN, existen técnicas más avanzadas que generan muestras sintéticas considerando densidad de datos, ruido y límites de clase, mejorando el rendimiento especialmente en la calase minoritaria. Entre estas se encuentras métodos como ASMOTE-ENN, que combina sobremuestreo adaptativo con limpieza de datos, o SOUP, que prioriza instancias "seguras" y elimina datos problemáticos mediante submuestreo [32]. Además, existen técnicas basadas en agrupación, como OOSI, y métodos híbridos que combinan sobremuestreo con submuestreo y aprendizaje ensemble, mostrando gran potencial para mejorar el balance de clases [33].

9 Análisis de impacto

9.1 Impacto personal

La realización del TFG ha representado una experiencia muy enriquecedora que ha favorecido mi desarrollo tanto técnico como personal. Afrontar la complejidad de trabajar con datos reales me permitió comprender la importancia de una preparación y limpieza exhaustiva, a pesar de que este proceso pueda resultar largo y meticuloso. Este aprendizaje fomentó en mí la paciencia y el compromiso necesarios para construir una base sólida en cualquier proyecto.

La aplicación de técnicas de procesamiento del lenguaje natural, como Word2Vec, me ayudó a comprender la dificultad que implica traducir el lenguaje humano a formatos comprensibles por las máquinas. Además, la gestión del desequilibrio de clases y la evaluación rigurosa de los modelos, especialmente al identificar errores como la fuga de datos, o el sobreajuste, reforzaron la necesidad de mantener un enfoque metodológico cuidadoso y crítico. Esta experiencia ha fortalecido mi capacidad analítica y mi confianza para afrontar problemas reales, al mismo tiempo que me ha enseñado a ser más reflexiva, constante y rigurosa en el trabajo académico y profesional.

9.2 Impacto social

Este estudio ofrece una visión relevante sobre cómo se perciben y comunican cultural y socialmente las drogas en el entorno digital. A través de Twitter, se identificaron patrones de lenguaje y asociaciones que pueden ser útiles para diseñar estrategias de prevención más eficaces y culturalmente adaptadas. Por ejemplo, la vinculación del éxtasis con connotaciones y del GHB con referencias, puede servir como base para comprender mejor las actitudes sociales hacia estas sustancias.

Este tipo de análisis también contribuye a una mayor comprensión de las comunidades digitales que existen hoy en día y sus formas de comunicación. No obstante, es importante considerar las implicaciones éticas, ya que el uso de datos provenientes de redes sociales puede generar preocupaciones relacionadas con la privacidad si no se manejan de forma adecuada.

9.3 Impacto económico y empresarial

Aunque el objetivo principal del trabajo no es económico, los resultados obtenidos pueden tener implicaciones indirectas en este ámbito. El análisis de redes sociales se presenta como una herramienta eficaz y económica para obtener información sobre parones de consumo de drogas, lo cual puede ayudar a optimizar la asignación de recursos en programas de prevención y tratamiento. Además, los métodos y resultados del estudio pueden ser de interés para empresas dedicadas al análisis de datos y la vigilancia de medios digitales, ya que ofrecen información útil para el diseño de estrategias en salud pública, marketing social o investigación social aplicada.

9.4 Impacto medioambiental

El impacto ambiental directo es mínimo, limitado principalmente al consumo energético derivado del uso de recursos computacionales. De manera muy indirecta, podría considerarse una sociedad más informada sobre los riesgos del consumo de drogas y con mejores estrategias de prevención puede favorecer entornos más saludables y sostenibles, pero esta conexión es marginal en el contexto específico de este trabajo y no constituye uno de sus objetivos principales.

9.5 Impacto cultural

El estudio ofrece una perspectiva interesante sobre las percepciones culturales del consumo de drogas, especialmente en contextos digitales. Los resultados permiten observar cómo se forman, comparten y transforman las actitudes hacia distintas sustancias en plataformas como Twitter. Un aspecto destacado es el uso del *slang* o jerga, que funciona como un marcador de identidad dentro de ciertas subculturas digitales, reflejando tanto el tono como la pertenencia a determinados grupos. Asimismo, se ha observado una mayor frecuencia de término relacionados con efectos psicológicos y emocionales, lo cual indica un enfoque más centrado en la experiencia subjetiva que en los efectos fisiológicos del consumo.

9.6 Referencia a los Objetivos de Desarrollo Sostenible de la Agenda 2030

Los resultados obtenidos en este trabajo tienen una clara vinculación con el ODS 3: Salud y Bienestar, específicamente para la Meta 3.5: "Fortalecer la prevención y el tratamiento del abuso de sustancias…", que busca fortalecer la prevención y el tratamiento del abuso de sustancias. La clasificación de efectos observados en los tweets aporta información útil para identificar tendencias de consumo y posibles riesgos, contribuyendo a la elaboración de estrategias de prevención más eficaces. Además, el enfoque en los efectos psicológicos del consumo se relaciona con la Meta 3.4, centrada en promover la salud mental y el bienestar. En este sentido, el TFG proporciona una base de datos y un análisis que pueden resultar de utilidad para alcanzar algunas de las metas de la Agenda 2030.

9.7 Decisiones basadas en la consideración del impacto

Durante el desarrollo del trabajo se tomaron diversas decisiones metodológicas con el objetivo de asegurar la calidad y utilidad de los resultados, La adopción del enfoque CRISP-DM permitió seguir una metodología estructurada y rigurosa, que favoreció la validez del análisis. La integración de la herramienta MetaMap permitió vincular la jerga y el lenguaje informal con conceptos biomédicos estandarizados, añadiendo solidez al análisis. Asimismo, se abordaron adecuadamente problemas como el desequilibrio de clases, lo que contribuyó a mejorar la fiabilidad de modelos. Finalmente, se puso especial atención en la preparación de los datos y la evaluación de modelos, garantizando así que las conclusiones fueran sólidas y útiles para futuras investigaciones o intervenciones.

10 Bibliografía

[1] J. Hilliard, «The Influence Of Social Media On Teen Drug Use», Addiction Center. Accedido: 13 de mayo de 2025. [En línea]. Disponible en: https://www.addictioncenter.com/community/social-media-teen-drug-use/

[2] H.-W. Meng, S. Kath, D. Li, y Q. C. Nguyen, «National substance use patterns on Twitter», *PLoS ONE*, vol. 12, n.º 11, p. e0187691, nov. 2017, doi: 10.1371/journal.pone.0187691.

[3] C. Castillo-Toledo *et al.*, «Insights from the Twittersphere: a cross-sectional study of public perceptions, usage patterns, and geographical differences of tweets discussing cocaine», *Front. Psychiatry*, vol. 15, mar. 2024, doi: 10.3389/fpsyt.2024.1282026.

[4] S. Kovandzich, «Many Post on Social Media Under the Influence of Drugs – and Regret It», Taylor & Francis Newsroom. Accedido: 13 de mayo de 2025. [En línea]. Disponible en: https://newsroom.taylorandfrancisgroup.com/many-post-on-social-media-under-the-influence-of-drugs-and-regret-it/

[5] J. Tassone, P. Yan, M. Simpson, C. Mendhe, V. Mago, y S. Choudhury, «Utilizing Deep Learning to Identify Drug Use on Twitter Data», 8 de marzo de 2020, *arXiv*: arXiv:2003.11522. doi: 10.48550/arXiv.2003.11522.

[6] T. Nasralah, O. El-Gayar, y Y. Wang, «Social Media Text Mining Framework for Drug Abuse: Development and Validation Study With an Opioid Crisis Case Analysis», *J. Med. Internet Res.*, vol. 22, n.º 8, p. e18350, ago. 2020, doi: 10.2196/18350.

[7] A. Alnawas, H. Alharbi, y M. M. Hassoun Al-Jawad, «Understanding the Influence Impact of Social Media on Drug Addiction: A Novel Sentiment Analysis Approach Using Multi-Level User Engagement Data», *Iraqi J. Electr. Electron. Eng.*, vol. 21, n.º 2, pp. 43-53, dic. 2025, doi: 10.37917/ijeee.21.2.5.

[8] A. Rodríguez-González *et al.*, «Identifying Polarity in Tweets from an Imbalanced Dataset about Diseases and Vaccines Using a Meta-Model Based on Machine Learning Techniques», *Appl. Sci.*, vol. 10, n.º 24, p. 9019, dic. 2020, doi: 10.3390/app10249019.

[9] A. Fisher, M. M. Young, D. Payer, K. Pacheco, C. Dubeau, y V. Mago, «Automating Detection of Drug-Related Harms on Social Media: Machine Learning Framework», *J. Med. Internet Res.*, vol. 25, p. e43630, sep. 2023, doi: 10.2196/43630.

[10] P. Chapman *et al.*, «Step-by-step data mining guide».

[11] «the-crisp-dm-model-the-new-blueprint-for-data-mining-shearer-colin.pdf». Accedido: 20 de mayo de 2025. [En línea]. Disponible en: https://mineracaodedados.wordpress.com/wp-content/uploads/2012/04/the-crisp-dm-model-the-new-blueprint-for-data-mining-shearer-colin.pdf

[12] D. Pyle, «Data Preparation for Data Mining».

[13] A. Gouloumis Contreras, «Analysis and characterisation of recreational drug effects using public sources», masters, E.T.S. de Ingenieros Informáticos (UPM), 2021. Accedido: 28 de mayo de 2025. [En línea]. Disponible en: https://oa.upm.es/68755/

[14] O. Bodenreider, «The Unified Medical Language System (UMLS): integrating biomedical terminology», *Nucleic Acids Res.*, vol. 32, n.º Database issue, pp. D267-D270, ene. 2004, doi: 10.1093/nar/gkh061.

[15] «Jerga - Concepto, tipos, ejemplos y diferencias con dialecto», https://concepto.de/. Accedido: 16 de mayo de 2025. [En línea]. Disponible en: https://concepto.de/jerga/

[16] R. Shyu y C. Weng, «Enabling Semantic Topic Modeling on Twitter Using MetaMap», *AMIA Summits Transl. Sci. Proc.*, vol. 2024, pp. 670-678, may 2024.

[17] J. Pain, J. Levacher, A. Quinquenel, y A. Belz, «Analysis of Twitter Data for Postmarketing Surveillance in Pharmacovigilance», en *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, B. Han, A. Ritter, L. Derczynski, W. Xu, y T. Baldwin, Eds., Osaka, Japan: The COLING 2016 Organizing Committee, dic. 2016, pp. 94-101. Accedido: 16 de mayo de 2025. [En línea]. Disponible en: https://aclanthology.org/W16-3914/

[18] F. Moons y E. Vandervieren, «Measuring agreement among several raters classifying subjects into one-or-more (hierarchical) nominal categories. A generalisation of Fleiss' kappa», 22 de marzo de 2023, *arXiv*: arXiv:2303.12502. doi: 10.48550/arXiv.2303.12502.

[19] T. Mikolov, K. Chen, G. Corrado, y J. Dean, «Efficient Estimation of Word Representations in Vector Space», 7 de septiembre de 2013, *arXiv*: arXiv:1301.3781. doi: 10.48550/arXiv.1301.3781.

[20] «RandomizedSearchCV», scikit-learn. Accedido: 1 de junio de 2025. [En línea]. Disponible en: https://scikit-learn/stable/modules/generated/sklearn.model_selection.RandomizedSearchCV.html

[21] «Nested versus non-nested cross-validation», scikit-learn. Accedido: 25 de mayo de 2025. [En línea]. Disponible en: https://scikit-learn/stable/auto_examples/model_selection/plot_nested_cross_validation_iris.html

[22] L. Breiman, J. Friedman, R. A. Olshen, y C. J. Stone, *Classification and Regression Trees*. New York: Chapman and Hall/CRC, 2017. doi: 10.1201/9781315139470.

[23] L. Breiman, «Random Forests», *Mach. Learn.*, vol. 45, n.º 1, pp. 5-32, oct. 2001, doi: 10.1023/A:1010933404324.

[24] P. McCullagh y J. A. Nelder, *Generalized Linear Models*. Boston, MA: Springer US, 1989. doi: 10.1007/978-1-4899-3242-6.

[25] T. Chen y C. Guestrin, «XGBoost: A Scalable Tree Boosting System», en *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, en KDD '16. New York, NY, USA: Association for Computing Machinery, ago. 2016, pp. 785-794. doi: 10.1145/2939672.2939785.

[26] Y. Freund y R. E. Schapire, «A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting», *J Comput Syst Sci*, vol. 55, n.º 1, pp. 119-139, ago. 1997, doi: 10.1006/jcss.1997.1504.

[27] L. Breiman, «Bagging predictors», *Mach. Learn.*, vol. 24, n.º 2, pp. 123-140, ago. 1996, doi: 10.1007/BF00058655.

[28] G. E. Hinton, «Connectionist learning procedures», *Artif. Intell.*, vol. 40, n.º 1-3, pp. 185-234, sep. 1989, doi: 10.1016/0004-3702(89)90049-0.

[29] J. Devlin, M.-W. Chang, K. Lee, y K. Toutanova, «BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding», 24 de mayo de 2019, *arXiv*: arXiv:1810.04805. doi: 10.48550/arXiv.1810.04805.

[30] Z. R. K. Rostam y G. Kertész, «Fine-Tuning Large Language Models for Scientific Text Classification: A Comparative Study», 27 de noviembre de 2024, *arXiv*: arXiv:2412.00098. doi: 10.48550/arXiv.2412.00098.

[31] T. B. Brown *et al.*, «Language Models are Few-Shot Learners», 22 de julio de 2020, *arXiv*: arXiv:2005.14165. doi: 10.48550/arXiv.2005.14165.

[32] N. V. Chawla, K. W. Bowyer, L. O. Hall, y W. P. Kegelmeyer, «SMOTE: Synthetic Minority Over-sampling Technique», *J. Artif. Intell. Res.*, vol. 16, pp. 321-357, jun. 2002, doi: 10.1613/jair.953.

[33] Y. Deng y M. Li, «An Adaptive and Robust Method for Oriented Oversampling With Spatial Information for Imbalanced Noisy Datasets», *IEEE Access*, vol. 11, pp. 122610-122624, 2023, doi: 10.1109/ACCESS.2023.3329560.

Anexo

Índice

Anexo I: Esquema previo del modelo Entidad-Relación 53

Anexo II: Script generado por MySQL Workbench para crear las tablas de la base de datos 54

Anexo III: Figuras complementarias en el análisis exploratorio 61

Anexo IV: Consulta SQL para crear el *dataset* final 74

Anexo V: Modelos y búsqueda de hiperparámetros 76

Anexo VI: Resultados de validación cruzada para las estrategias de entrenamiento evaluadas 77

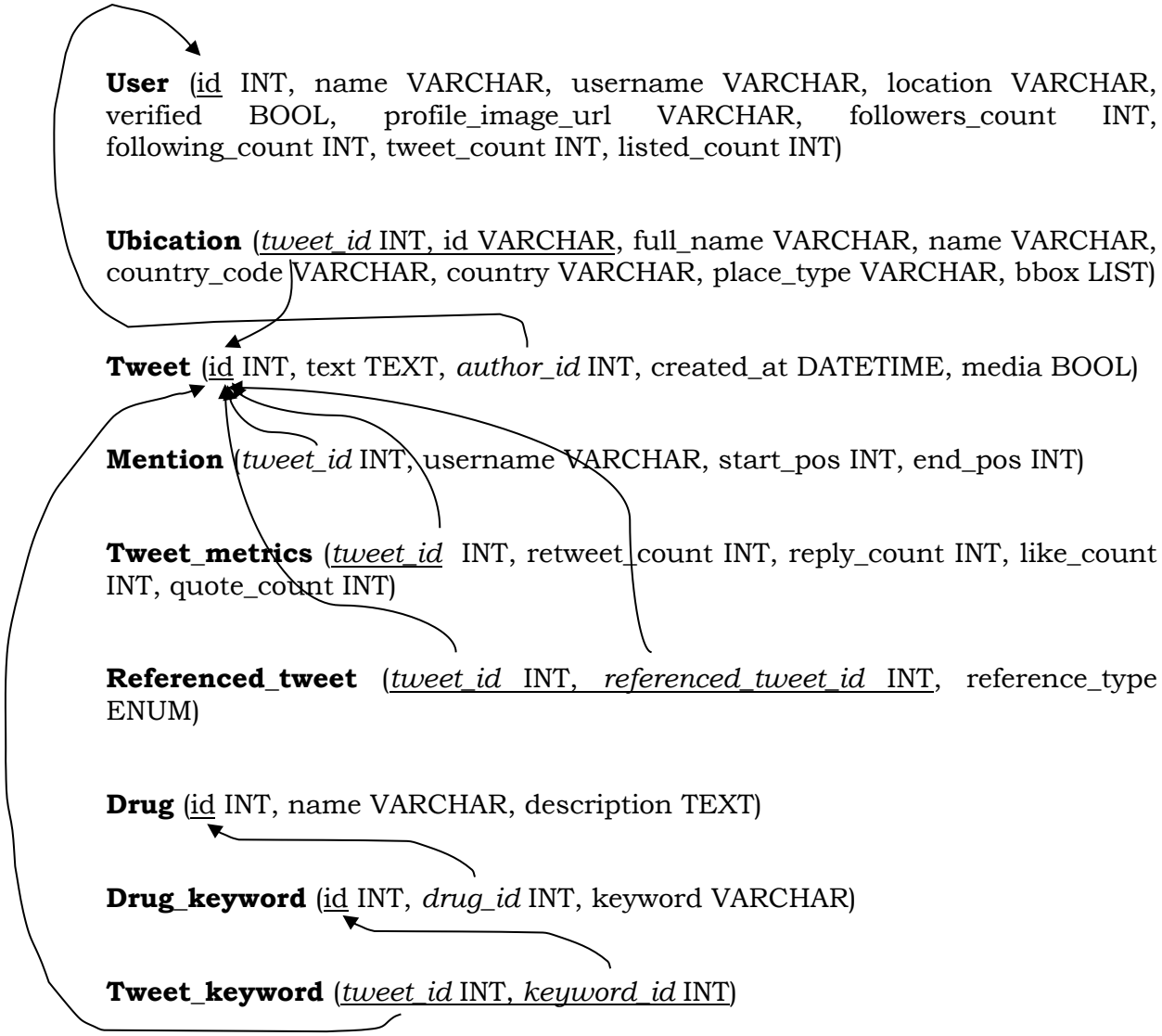
Índice de figuras del anexo

Figura A.1. Distribución de tweets por keyword y por droga. 61
Figura A.2. Distribución de tweets por año y por droga. 62
Figura A.3. Distribución de tweets por mes y por droga. 63
Figura A.4. Distribución de tweets por día y por droga. 64
Figura A.5. Distribución de tweets por país y por droga. 65
Figura A.6. Distribución de tweets por tipo de referencia. 66
Figura A.7. Distribución de engagement por tipo de usuario y número de seguidores. 67
Figura A.8. Promedio de engagement por presencia de media. 67
Figura A.9. Distribución del engagement por día. 68
Figura A.10. Proporción de tweets con slang. 68
Figura A.11. Distribución de anotaciones por slang. 69
Figura A.12. Distribución de tweets por slang. 69
Figura A.13. Distribución de tweets por usuario verificado y slang. 70
Figura A.14. Distribución de TUI por anotación. 70
Figura A.15. Distribución de tweets por TUI (menp) por droga. 71
Figura A.16. Distribución de tweets por MetaMap y por droga. 72
Figura A.17. WordCloud de MetaMap por droga. 73

Índice de tablas del anexo

Tabla A.1. Número de tweets por nombre de usuario no verificado. 66
Tabla A.2. Número de tweets por nombre de usuario verificado. 66
Tabla A.3. Tabla de métricas de validación cruzada para las cuatro estrategias de entrenamiento. 77

Anexo I: Esquema previo del modelo Entidad-Relación



En el esquema hemos seguido las siguientes convenciones para la representación de los atributos de cada entidad, además de los tipos de cada atributo:

- Las **entidades** se representan en **negrita** (por ejemplo, ***User***, ***Tweet***, ***Mention***).
- Las **claves primarias** (***Primary Keys* - PK**) están subrayadas para diferenciarlas del resto de atributos. Estas claves identifican de manera única cada registro dentro de una entidad.
- Las **claves foráneas** (***Foreign Keys* - FK**) están en *cursiva* para indicar que establecen una relación con otra entidad. Estas claves permiten la integridad referencial dentro del modelo.
- El **tipo de cada atributo** se especifica junto a su nombre, definiendo si es un entero (INT), una cadena de caracteres (VARCHAR), un valor booleano (BOOL), una fecha (DATETIME), entre otros.
- Las **flechas** indican la relación entre las entidades, mostrando qué clave foránea hace referencia a qué clave primaria en otra tabla.

Anexo II: Script generado por MySQL Workbench para crear las tablas de la base de datos

```
-- MySQL Workbench Forward Engineering


SET @OLD_UNIQUE_CHECKS=@@UNIQUE_CHECKS, UNIQUE_CHECKS=0;

SET @OLD_FOREIGN_KEY_CHECKS=@@FOREIGN_KEY_CHECKS, FOREIGN_KEY_CHECKS=0;

SET @OLD_SQL_MODE=@@SQL_MODE,
SQL_MODE='ONLY_FULL_GROUP_BY,STRICT_TRANS_TABLES,NO_ZERO_IN_DATE,NO_ZERO_
DATE,ERROR_FOR_DIVISION_BY_ZERO,NO_ENGINE_SUBSTITUTION';


-- -----------------------------------------------------

-- Schema twitter_analysis

-- -----------------------------------------------------

CREATE SCHEMA IF NOT EXISTS `twitter_analysis` DEFAULT CHARACTER SET utf8
;

USE `twitter_analysis` ;


-- -----------------------------------------------------

-- Table `twitter_analysis`.`user`

-- -----------------------------------------------------

CREATE TABLE IF NOT EXISTS `twitter_analysis`.`user` (

  `id` BIGINT NOT NULL,

  `name` VARCHAR(255) NULL,

  `username` VARCHAR(255) NULL,

  `location` VARCHAR(255) NULL,

  `verified` TINYINT NULL,

  `profile_image_url` VARCHAR(255) NULL,

  `followers_count` INT NULL,

  `following_count` INT NULL,

  `tweet_count` INT NULL,

  `listed_count` INT NULL,

  PRIMARY KEY (`id`))

ENGINE = InnoDB;


-- -----------------------------------------------------
```

```
-- Table `twitter_analysis`.`tweet`
-- -----------------------------------------------------
CREATE TABLE IF NOT EXISTS `twitter_analysis`.`tweet` (
  `id` BIGINT NOT NULL,
  `text` TEXT NULL,
  `author_id` BIGINT  NULL,
  `created_at` DATETIME NULL,
  `media` TINYINT NULL,
  PRIMARY KEY (`id`),
    FOREIGN KEY (`author_id`)
    REFERENCES `twitter_analysis`.`user` (`id`)
    ON DELETE RESTRICT
    ON UPDATE CASCADE)
ENGINE = InnoDB;


-- -----------------------------------------------------
-- Table `twitter_analysis`.`tweet_metrics`
-- -----------------------------------------------------
CREATE TABLE IF NOT EXISTS `twitter_analysis`.`tweet_metrics` (
  `tweet_id` BIGINT NOT NULL,
  `retweet_count` INT NULL,
  `reply_count` INT NULL,
  `like_count` INT NULL,
  `quote_count` INT NULL
  PRIMARY KEY (`tweet_id`),
    FOREIGN KEY (`tweet_id`)
    REFERENCES `twitter_analysis`.`tweet` (`id`)
    ON DELETE RESTRICT
    ON UPDATE CASCADE)
ENGINE = InnoDB;


-- -----------------------------------------------------
-- Table `twitter_analysis`.`mention`
-- -----------------------------------------------------
```

```
CREATE TABLE IF NOT EXISTS `twitter_analysis`.`mention` (
  `tweet_id` BIGINT NOT NULL,
  `username` VARCHAR(255) NULL,
  `start_pos` INT NULL,
  `end_pos` INT NULL,
  PRIMARY KEY (`tweet_id`),
    FOREIGN KEY (`tweet_id`)
    REFERENCES `twitter_analysis`.`tweet` (`id`)
    ON DELETE RESTRICT
    ON UPDATE CASCADE)
ENGINE = InnoDB;


-- -----------------------------------------------------
-- Table `twitter_analysis`.`referenced_tweet`
-- -----------------------------------------------------
CREATE TABLE IF NOT EXISTS `twitter_analysis`.`referenced_tweet` (
  `tweet_id` BIGINT NOT NULL,
  `referenced_tweet_id` BIGINT NOT NULL,
  `reference_type` ENUM('replied_to', 'quoted') NULL,
  PRIMARY KEY (`referenced_tweet_id`, `tweet_id`))
ENGINE = InnoDB;


-- -----------------------------------------------------
-- Table `twitter_analysis`.`drug`
-- -----------------------------------------------------
CREATE TABLE IF NOT EXISTS `twitter_analysis`.`drug` (
  `id` INT NOT NULL,
  `name` VARCHAR(100) NULL,
  `description` TEXT NULL,
  PRIMARY KEY (`id`))
ENGINE = InnoDB;


-- -----------------------------------------------------
-- Table `twitter_analysis`.`drug_keyword`
```

```
-- -----------------------------------------------------
CREATE TABLE IF NOT EXISTS `twitter_analysis`.`drug_keyword` (
  `id` INT NOT NULL,
  `drug_id` INT NULL,
  `keyword` VARCHAR(100) NULL,
  PRIMARY KEY (`id`),
    FOREIGN KEY (`drug_id`)
    REFERENCES `twitter_analysis`.`drug` (`id`)
    ON DELETE RESTRICT
    ON UPDATE CASCADE)
ENGINE = InnoDB;


-- -----------------------------------------------------
-- Table `twitter_analysis`.`tweet_keyword`
-- -----------------------------------------------------
CREATE TABLE IF NOT EXISTS `twitter_analysis`.`tweet_keyword` (
  `tweet_id` BIGINT NOT NULL,
  `keyword_id` INT NOT NULL,
  PRIMARY KEY (`tweet_id`, `keyword_id`),
    FOREIGN KEY (`tweet_id`)
    REFERENCES `twitter_analysis`.`tweet` (`id`)
    ON DELETE RESTRICT
    ON UPDATE CASCADE,
    FOREIGN KEY (`keyword_id`)
    REFERENCES `twitter_analysis`.`drug_keyword` (`id`)
    ON DELETE RESTRICT
    ON UPDATE CASCADE)
ENGINE = InnoDB;


-- -----------------------------------------------------
-- Table `twitter_analysis`.`ubication`
-- -----------------------------------------------------
CREATE TABLE IF NOT EXISTS `twitter_analysis`.`ubication` (
  `tweet_id` BIGINT NOT NULL,
```

```
  `id` VARCHAR(255) NULL,
  `name` VARCHAR(255) NULL,
  `full_name` VARCHAR(255) NULL,
  `country_code` CHAR(2) NULL,
  `country` VARCHAR(255) NULL,
  `place_type` VARCHAR(255) NULL,
  `bbox` JSON NULL,
  PRIMARY KEY (`tweet_id`),
    FOREIGN KEY (`tweet_id`)
    REFERENCES `twitter_analysis`.`tweet` (`id`)
    ON DELETE RESTRICT
    ON UPDATE CASCADE)
ENGINE = InnoDB;

CREATE TABLE IF NOT EXISTS `twitter_analysis`.`slang` (
  `id` INT NOT NULL,
  `concept` VARCHAR(255) NULL,
  `annotation1` VARCHAR(10) NULL,
  `annotation2` VARCHAR(10) NULL,
  `annotation3` VARCHAR(10) NULL,
  PRIMARY KEY (`id`))
ENGINE = InnoDB;

CREATE TABLE IF NOT EXISTS `twitter_analysis`.`metamap` (
  `cui` CHAR(8) NOT NULL,
  `term` VARCHAR(255) NULL,
  `tui` CHAR(4) NULL,
  `annotation1` VARCHAR(10) NULL,
  `annotation2` VARCHAR(10) NULL,
  `annotation3` VARCHAR(10) NULL,
  PRIMARY KEY (`cui`),
   FOREIGN KEY (`tui`)
    REFERENCES `twitter_analysis`.`semantic_type` (`tui`)
    ON DELETE RESTRICT
```

```
    ON UPDATE CASCADE)
ENGINE = InnoDB;

CREATE TABLE IF NOT EXISTS `twitter_analysis`.`semantic_type` (
  `tui` CHAR(4) NOT NULL,
  `abbreviation` CHAR(4) NULL,
  `description` VARCHAR(50) NULL,
  `annotation` VARCHAR(45) NULL,
  PRIMARY KEY (`tui`))
ENGINE = InnoDB;

CREATE TABLE IF NOT EXISTS `twitter_analysis`.`tweet_metamap` (
  `tweet_id` BIGINT(20) NOT NULL,
  `cui` CHAR(8) NOT NULL,
  `tui`CHAR(4) NOT NULL,
  `matched_words` TEXT NULL,
  `positional_info` TEXT NULL,
  PRIMARY KEY (`cui`, `tweet_id`, `tui`),
    FOREIGN KEY (`tweet_id`)
      REFERENCES `twitter_analysis`.`tweet` (`id`)
      ON DELETE RESTRICT
      ON UPDATE CASCADE,
    FOREIGN KEY (`cui`)
      REFERENCES `twitter_analysis`.`metamap` (`cui`)
      ON DELETE RESTRICT
      ON UPDATE CASCADE,
  FOREIGN KEY (`tui`)
    REFERENCES `twitter_analysis`.`semantic_type` (`tui`))
ENGINE = InnoDB;

CREATE TABLE IF NOT EXISTS `twitter_analysis`.`tweet_slang` (
  `tweet_id` bigint NOT NULL,
  `slang_id` INT NOT NULL,
  PRIMARY KEY (`tweet_id`, `slang_id`),
```

```
    FOREIGN KEY (`tweet_id`)
    REFERENCES `twitter_analysis`.`tweet` (`id`)
    ON DELETE RESTRICT
    ON UPDATE CASCADE,
    FOREIGN KEY (`slang_id`)
    REFERENCES `twitter_analysis`.`slang` (`id`)
    ON DELETE RESTRICT
    ON UPDATE CASCADE)
ENGINE = InnoDB;


SET SQL_MODE=@OLD_SQL_MODE;
SET FOREIGN_KEY_CHECKS=@OLD_FOREIGN_KEY_CHECKS;
SET UNIQUE_CHECKS=@OLD_UNIQUE_CHECKS;
```

Anexo III: Figuras complementarias en el análisis exploratorio

La **Figura A.1** muestra las 10 *keywords* con mayor número de tweets para cada droga.

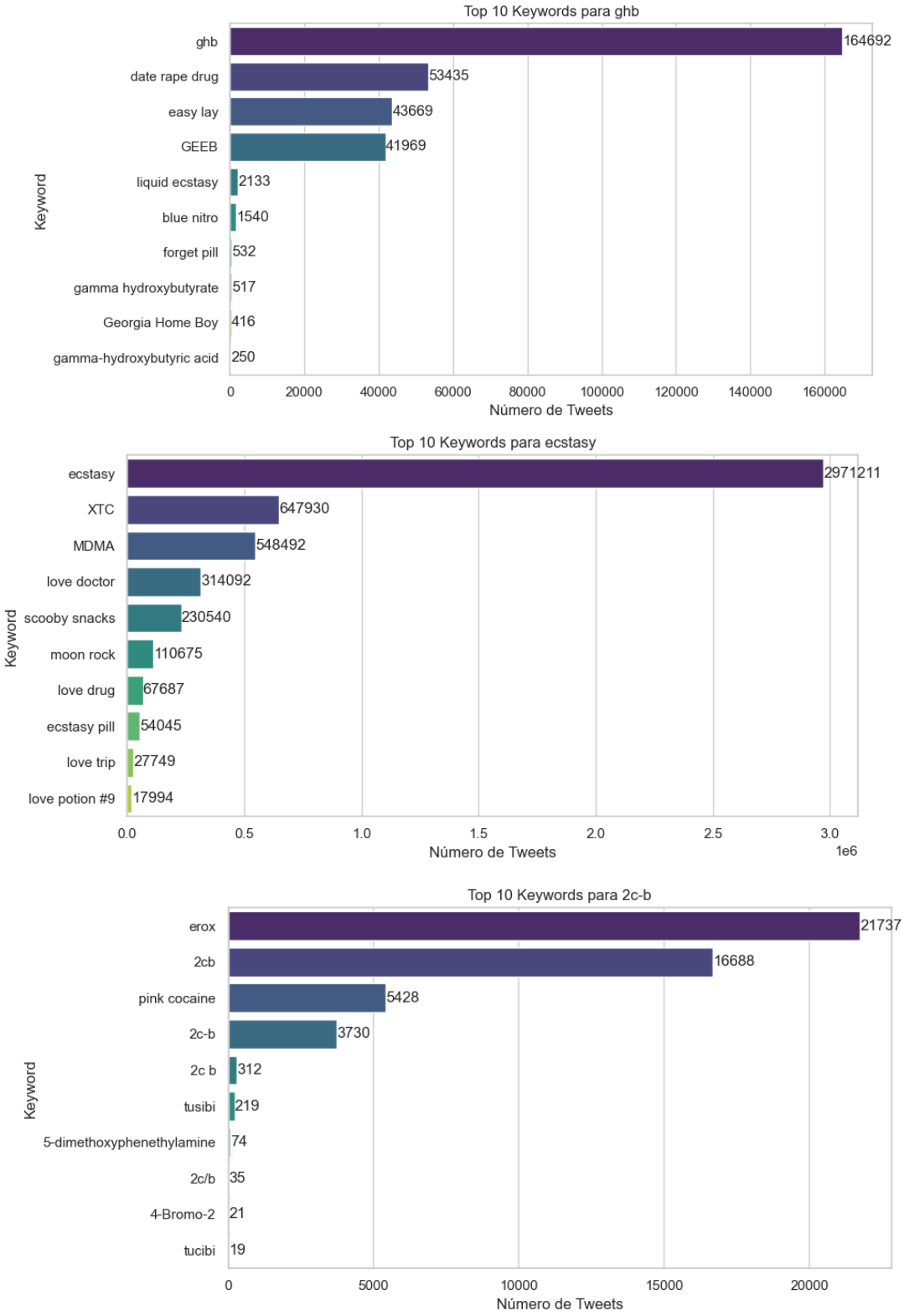


Figura A.1. Distribución de tweets por keyword y por droga.

La evolución del número de tweets publicados por droga a lo largo de los años se muestra en el **Figura A.2**.

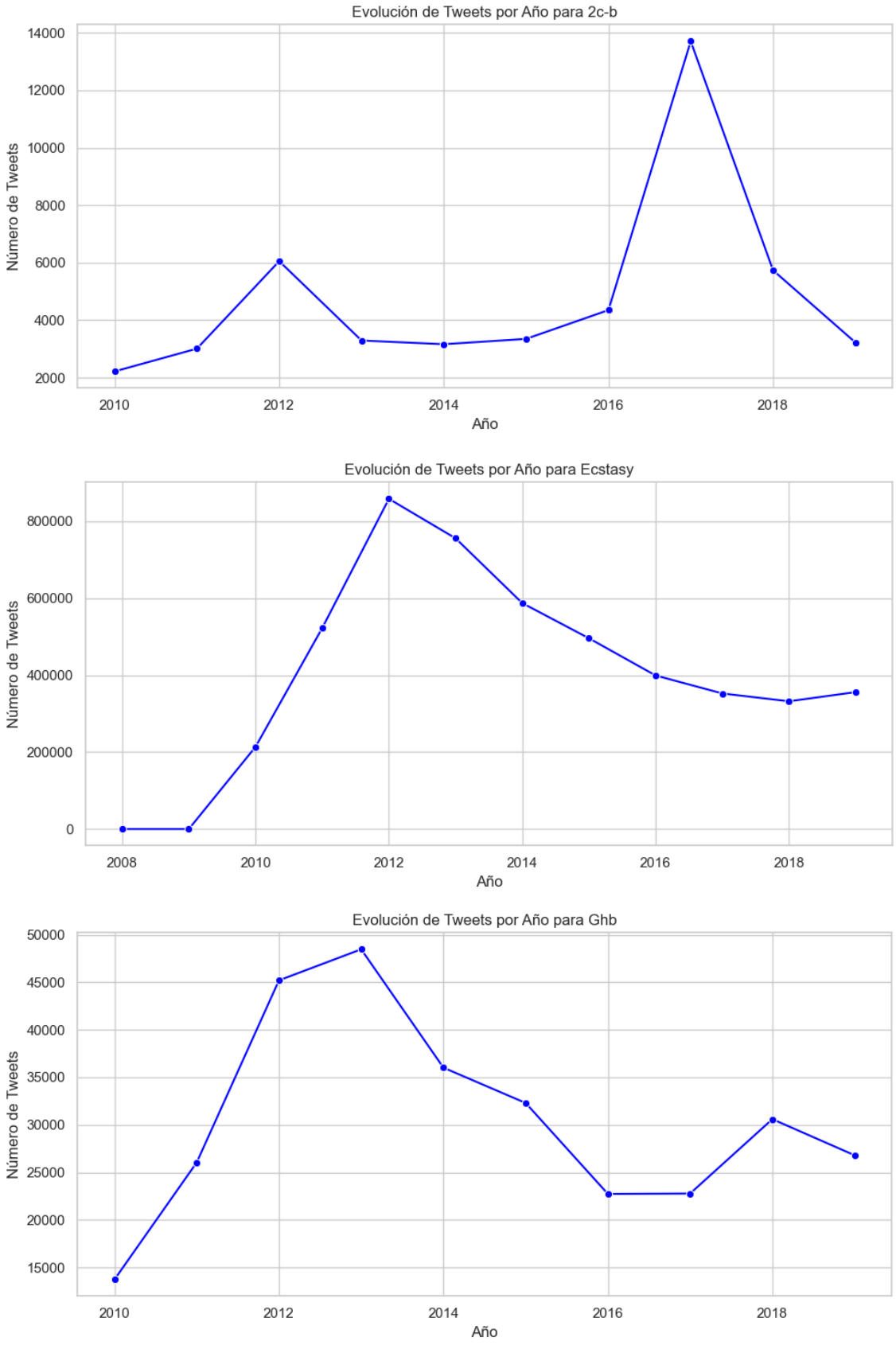


Figura A.2. Distribución de tweets por año y por droga.

El análisis mensual que se puede ver en la **Figura A.3**, muestra el número de tweets publicados cada mes, por droga.

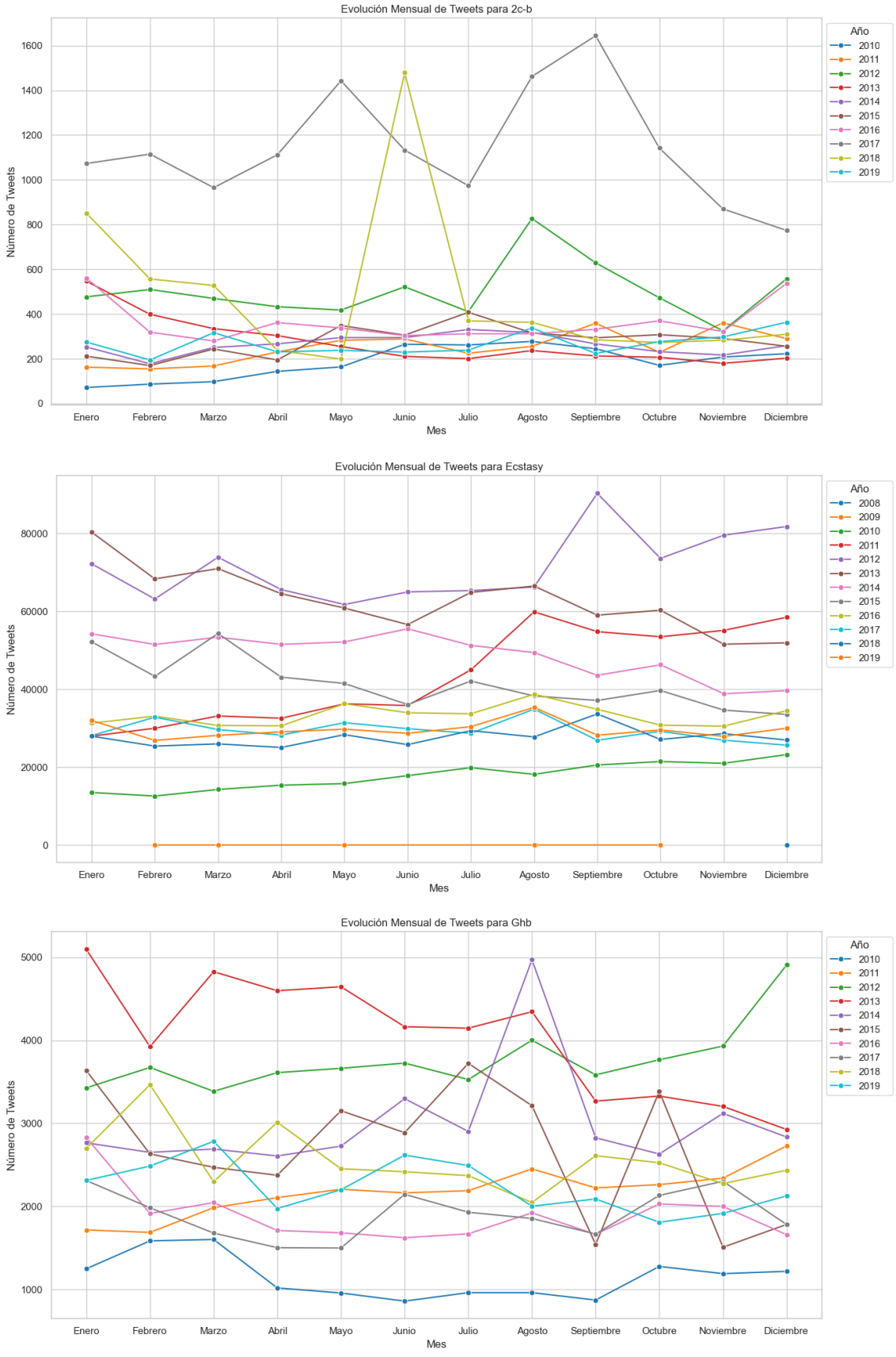


Figura A.3. Distribución de tweets por mes y por droga.

A continuación, se presenta en la **Figura A.4**, el número de tweets publicados a lo largo de un día por droga.

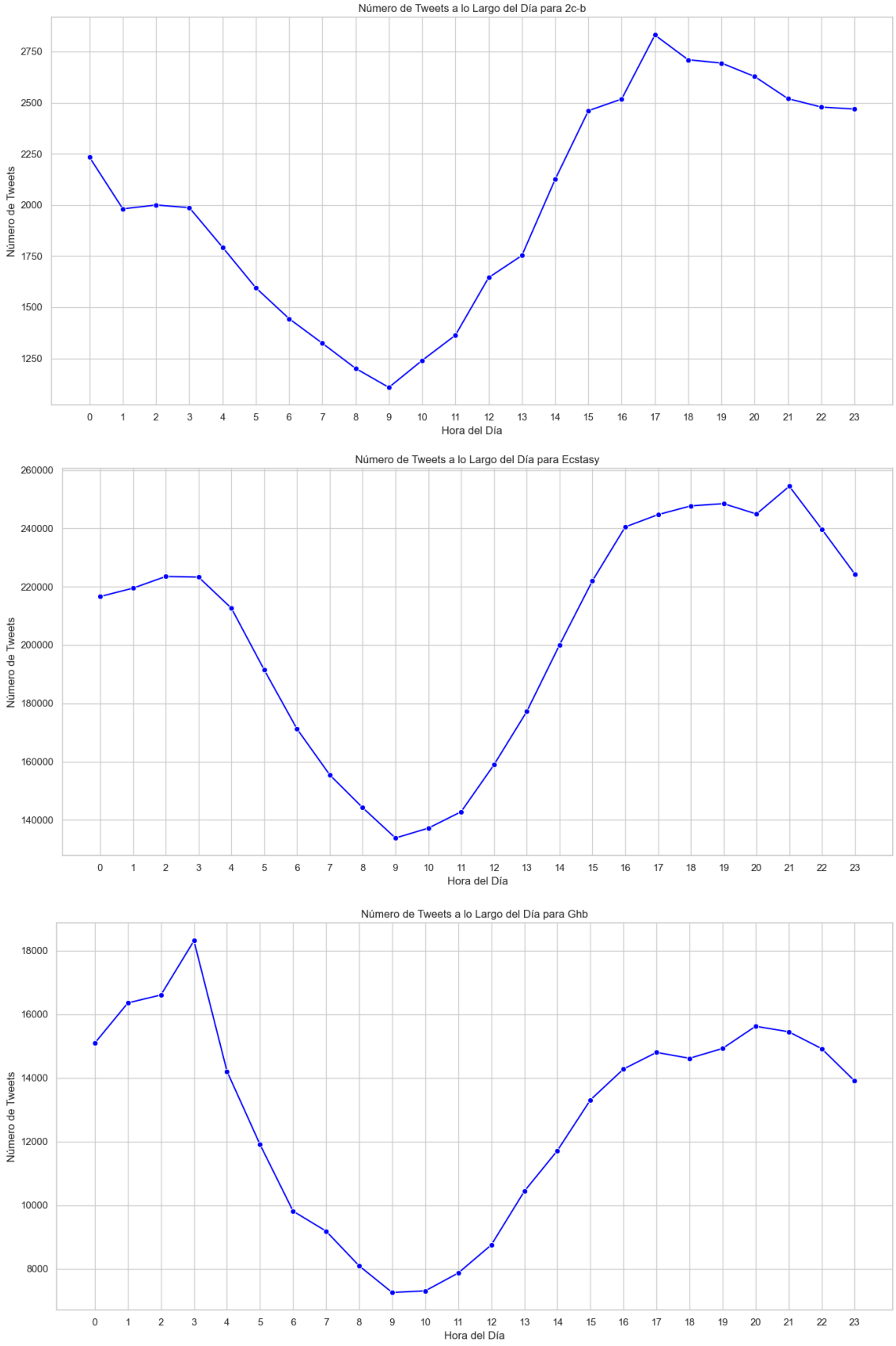


Figura A.4. Distribución de tweets por día y por droga.

La **Figura A.5** muestra los 10 primeros países que han publicado más tweets, sobre cada una de las drogas analizadas.

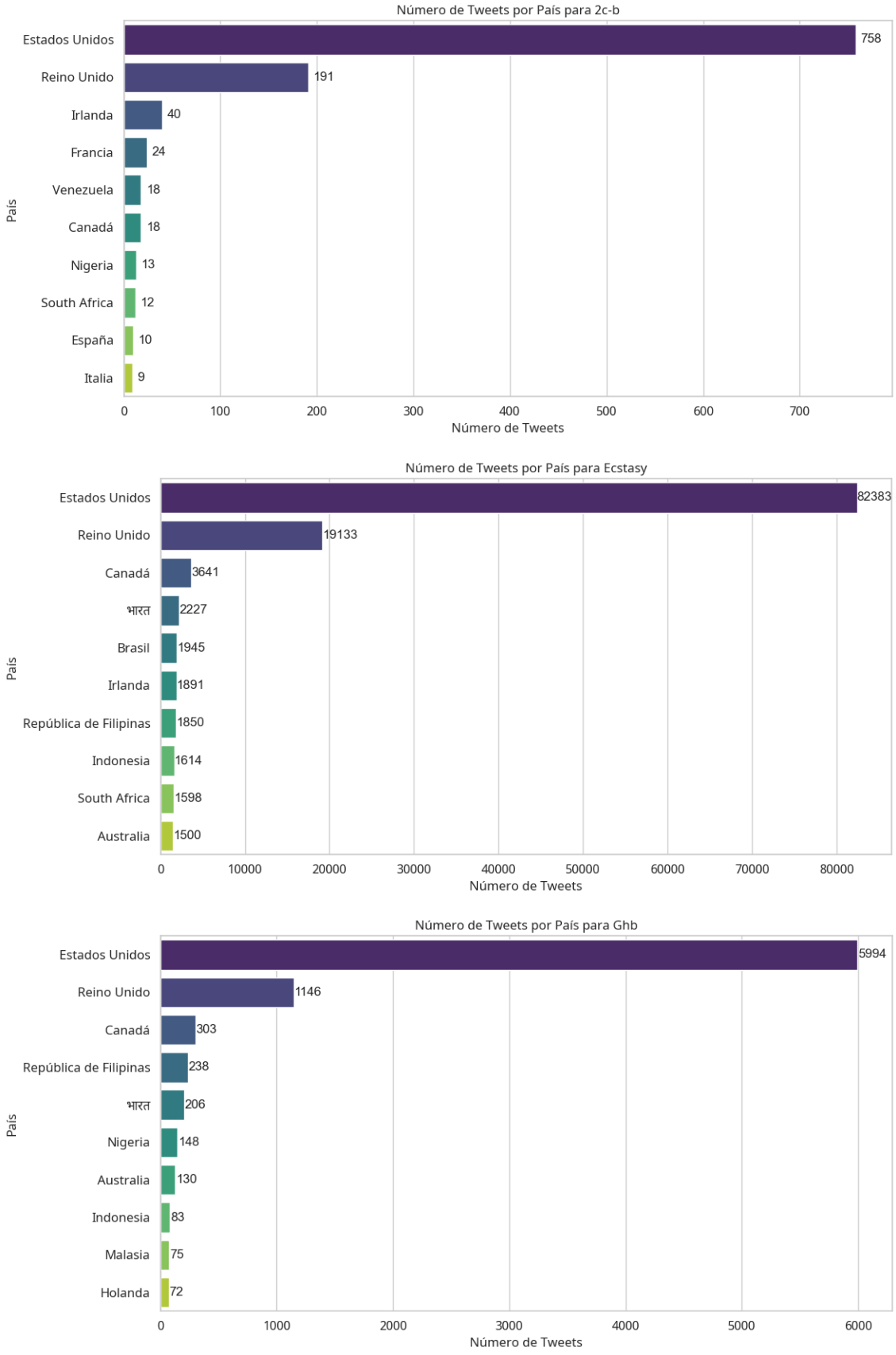


Figura A.5. Distribución de tweets por país y por droga.

La **Tabla A.1** presenta el número total de tweets generados por usuarios no verificados destacando que el usuario 'Kamasustra_neue' es el más activo con 25.070 tweets.

Tabla A.1. Número de tweets por nombre de usuario no verificado.

Nombre de usuario no verificado	Nº total de tweets
Kamasutra_neue	25.070
ASSENTCitizens	11.313
redditDrugs	6.380
djxtcnet	5.722
codepoetrybot	5.653

En la **Tabla A.2** se muestra el número total de tweets de usuarios verificados, donde 'urbandictionary' lidera con 8.618 tweets.

Tabla A.2. Número de tweets por nombre de usuario verificado.

Nombre de usuario verificado	Nº total de tweets
urbandictionary	8.618
MAPS	2.182
JeremyDuns	1.218
DoubleJPlays	840
jconnollybooks	749

La **Figura A.6** ilustra la distribución de tweets según el tipo de referencia en los tweets, destacando que el tipo de referencia más común es 'replied_to'.

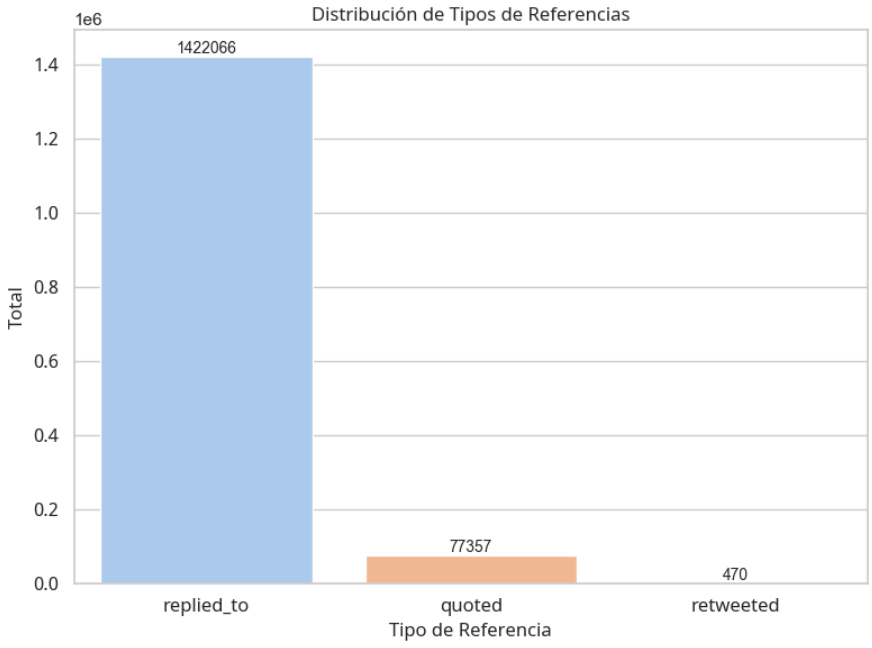


Figura A.6. Distribución de tweets por tipo de referencia.

Se puede observar en la **Figura A.7** cómo varía el *engagement* en función del tipo de usuario y el número de seguidores, mostrando diferencias claras entre usuarios verificados y no verificados.

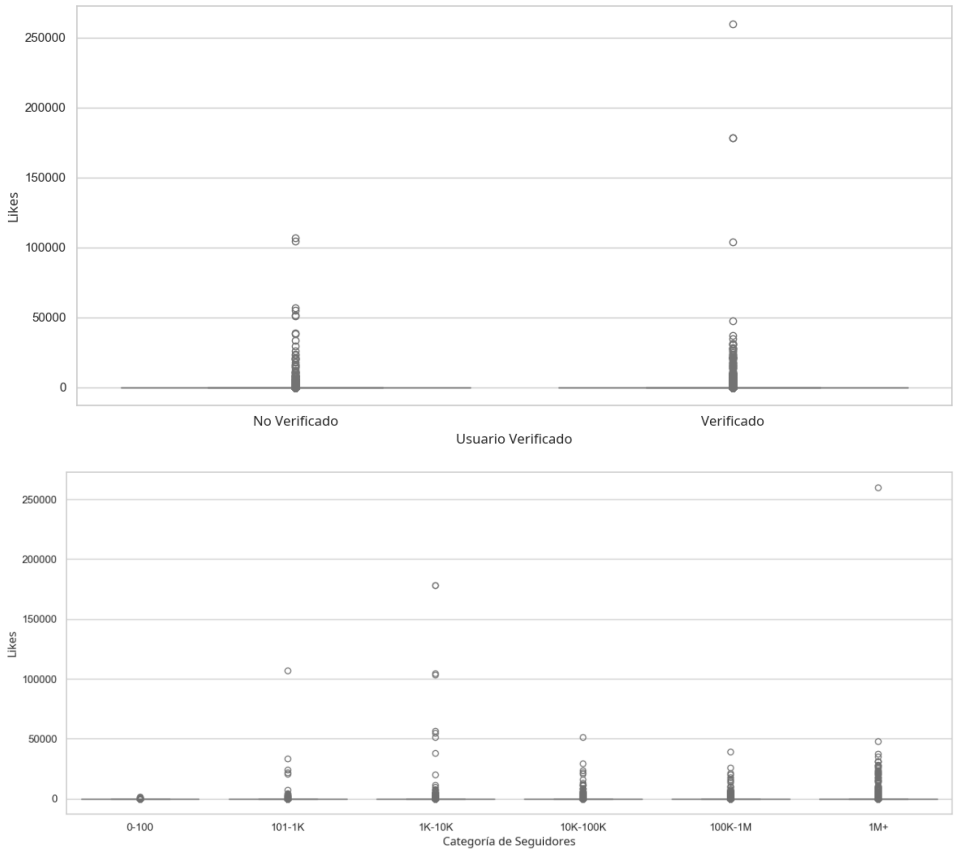


Figura A.7. Distribución de engagement por tipo de usuario y número de seguidores.

La influencia de la presencia de contenido multimedia en los tweets se refleja en la **Figura A.8**, donde los tweets con media presentan un nivel de interacción superior.

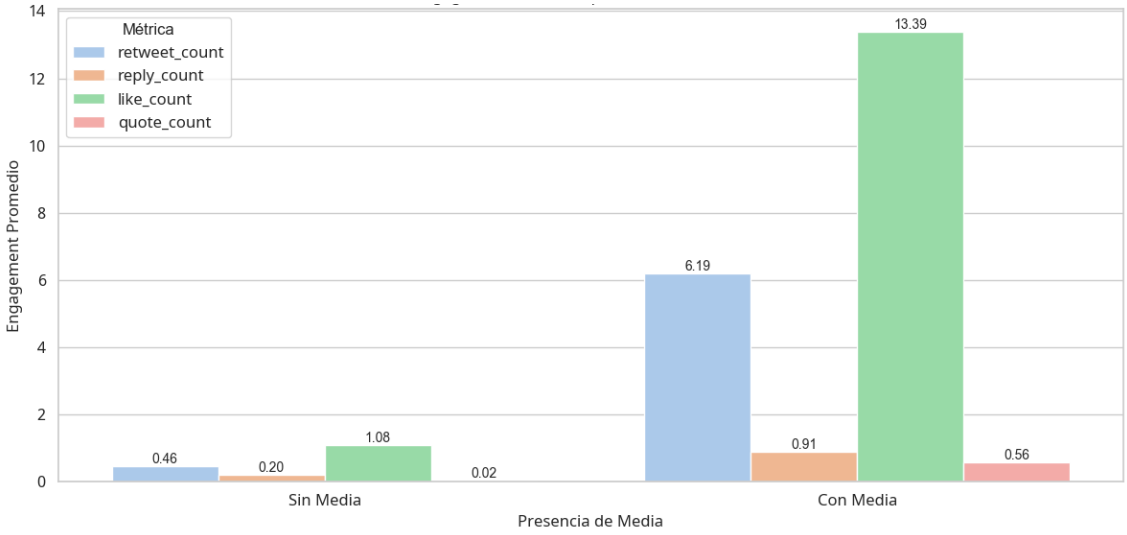


Figura A.8. Promedio de engagement por presencia de media.

En la **Figura A.9** se muestra el *engagement* promedio de métricas a lo largo de un día, está claro que el número de retweets es la interacción más utilizada entre los usuarios, además la hora punta donde se registran más interacciones es a las 11:00 de la mañana.

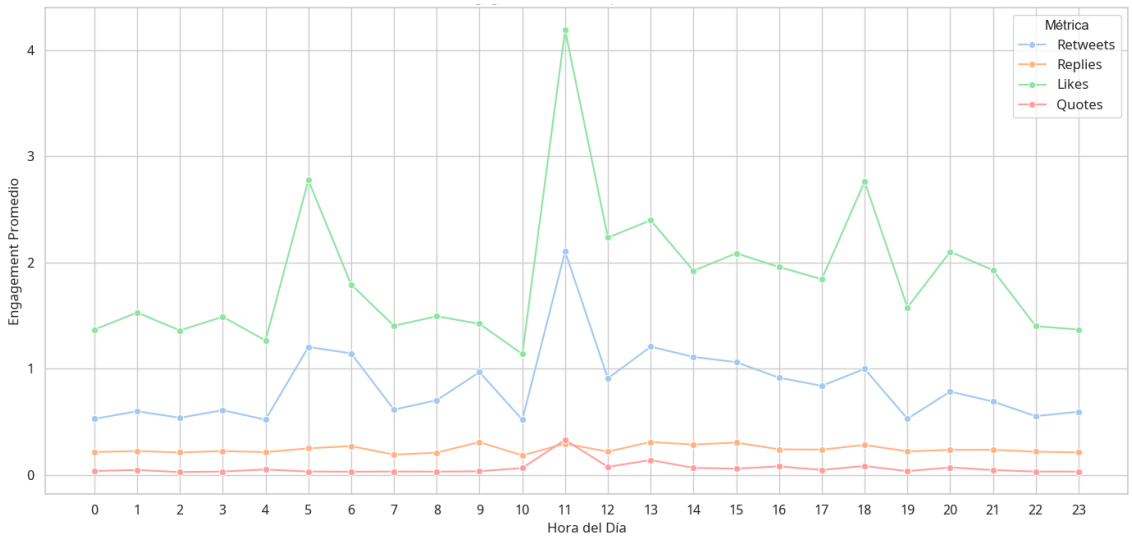


Figura A.9. Distribución del engagement por día.

En cuanto al uso de lenguaje informal, la **Figura A.10** muestra la proporción de tweets que contienen *slang*.

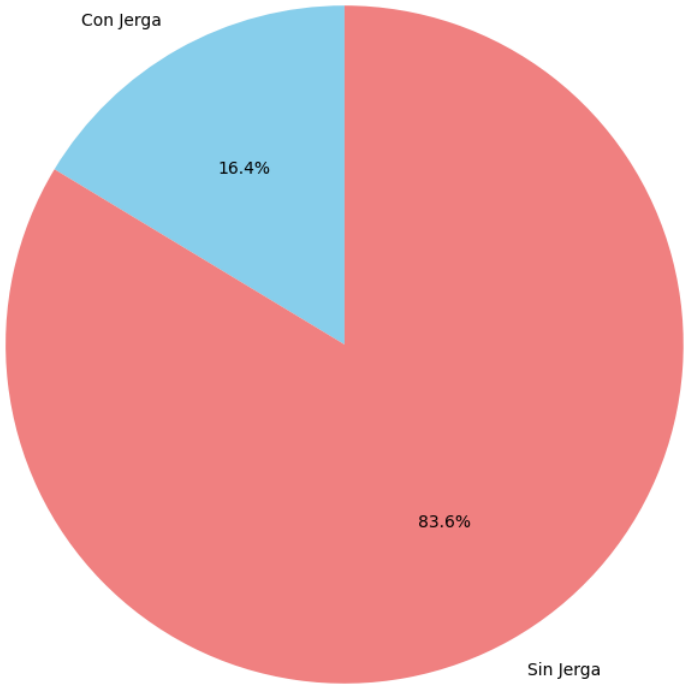


Figura A.10. Proporción de tweets con slang.

La distribución y frecuencia de las anotaciones relacionadas con *slang* se detallan en la **Figura A.11**.

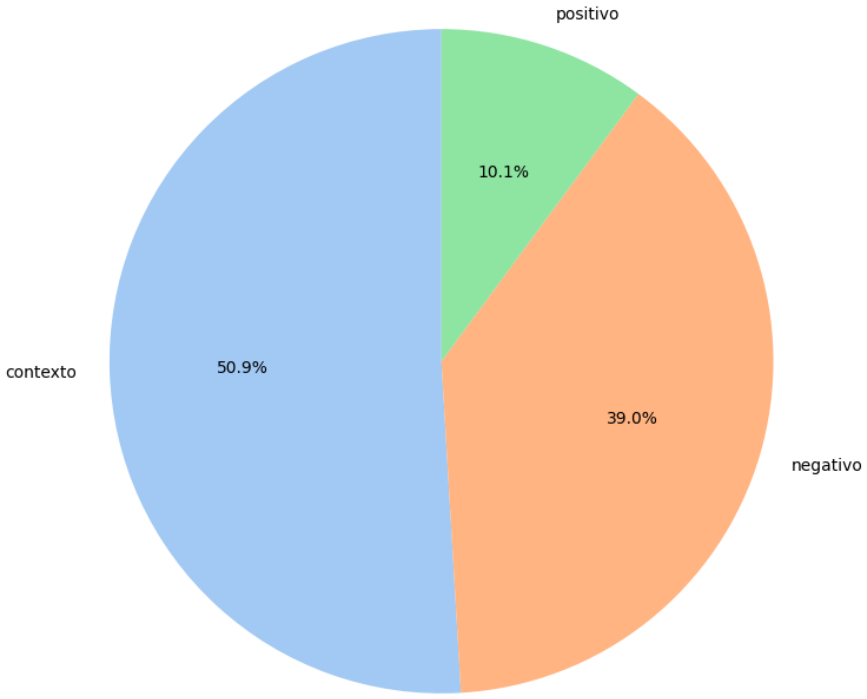


Figura A.11. Distribución de anotaciones por slang.

Las 10 *slangs* más utilizadas por los usuarios se pueden ver en la **Figura A.12**.

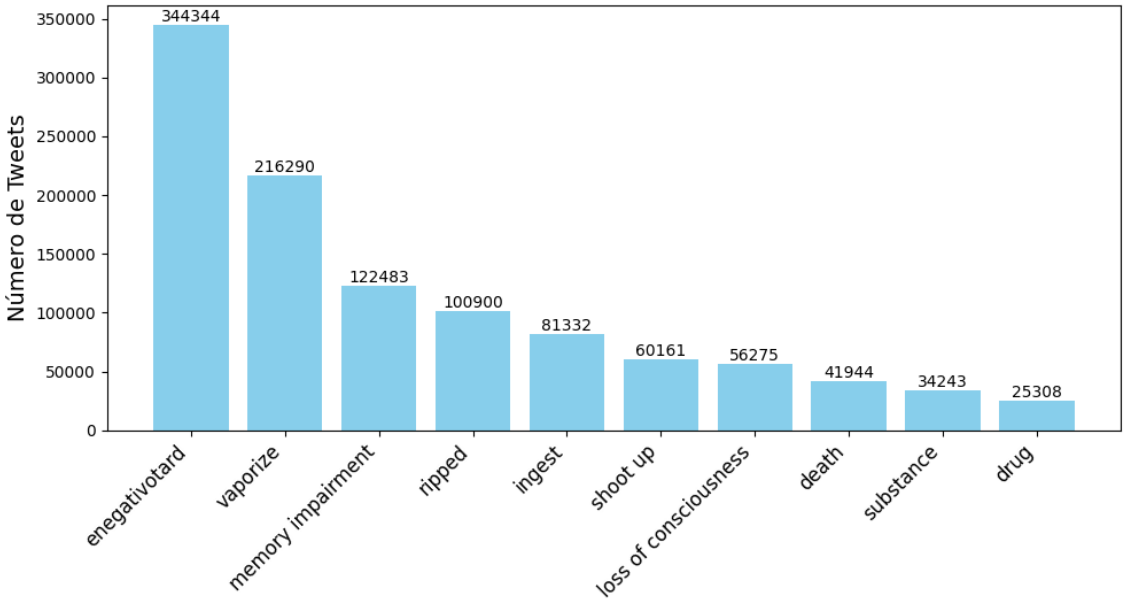


Figura A.12. Distribución de tweets por slang.

La distribución de tweets por usuario verificado y uso de *slang* la podemos ver en la **Figura A.13**, es evidente que la mayoría de tweets que mencionan slang son publicados por usuarios no verificados.

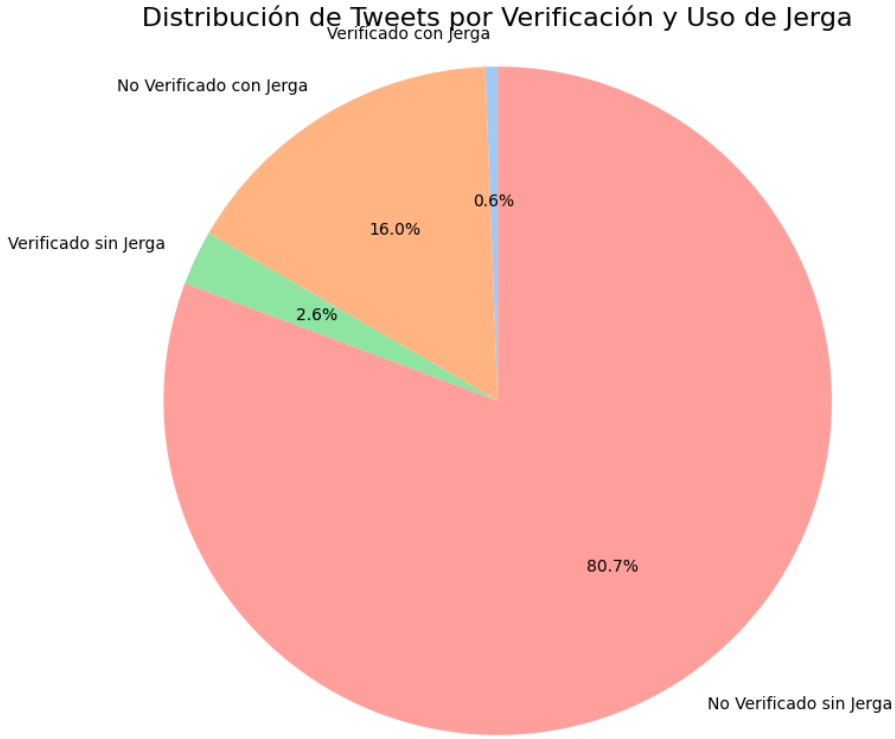


Figura A.13. Distribución de tweets por usuario verificado y slang.

La **Figura A.14** exhibe la distribución de TUIs de acuerdo con las anotaciones.

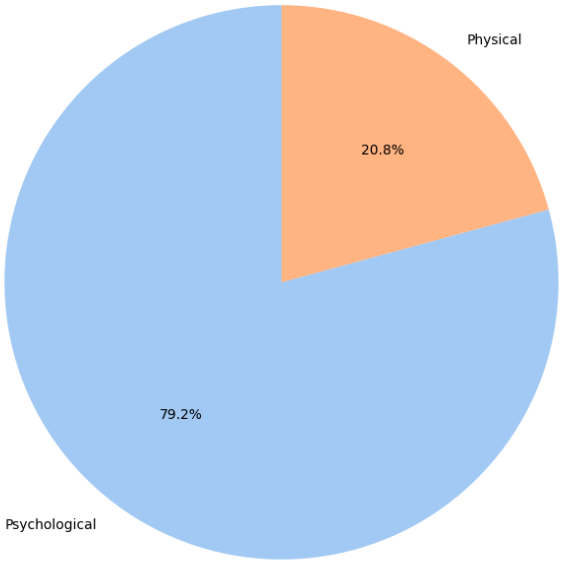


Figura A.14.Distribución de TUI por anotación.

La **Figura A.15** detalla la distribución del TUI (menp) en tweets relacionados con las distintas, señalando diferencias en el uso del lenguaje.

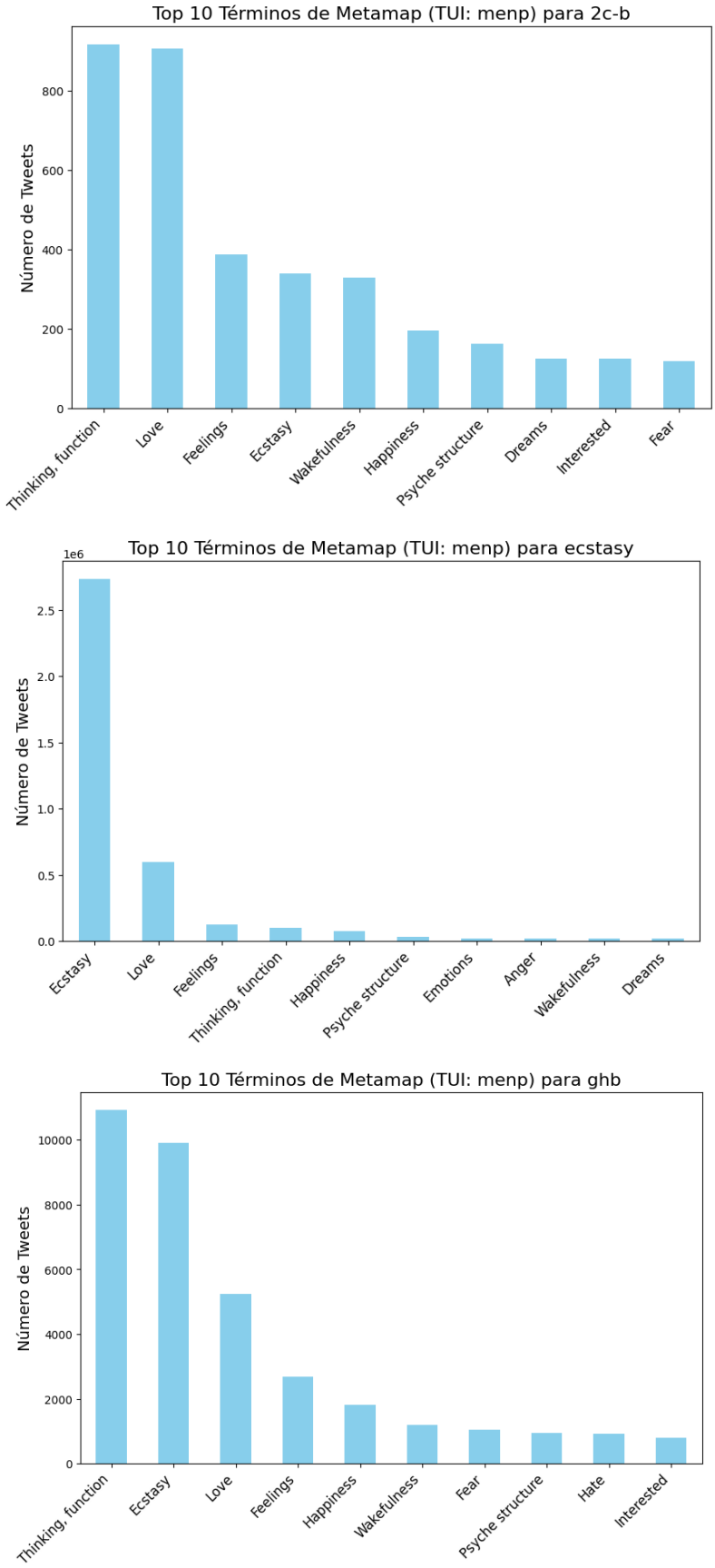


Figura A.15. Distribución de tweets por TUI (menp) por droga.

A continuación, en la **Figura A.16** se muestran los 10 términos de MetaMap que aparecen en mayor número de tweets por droga.

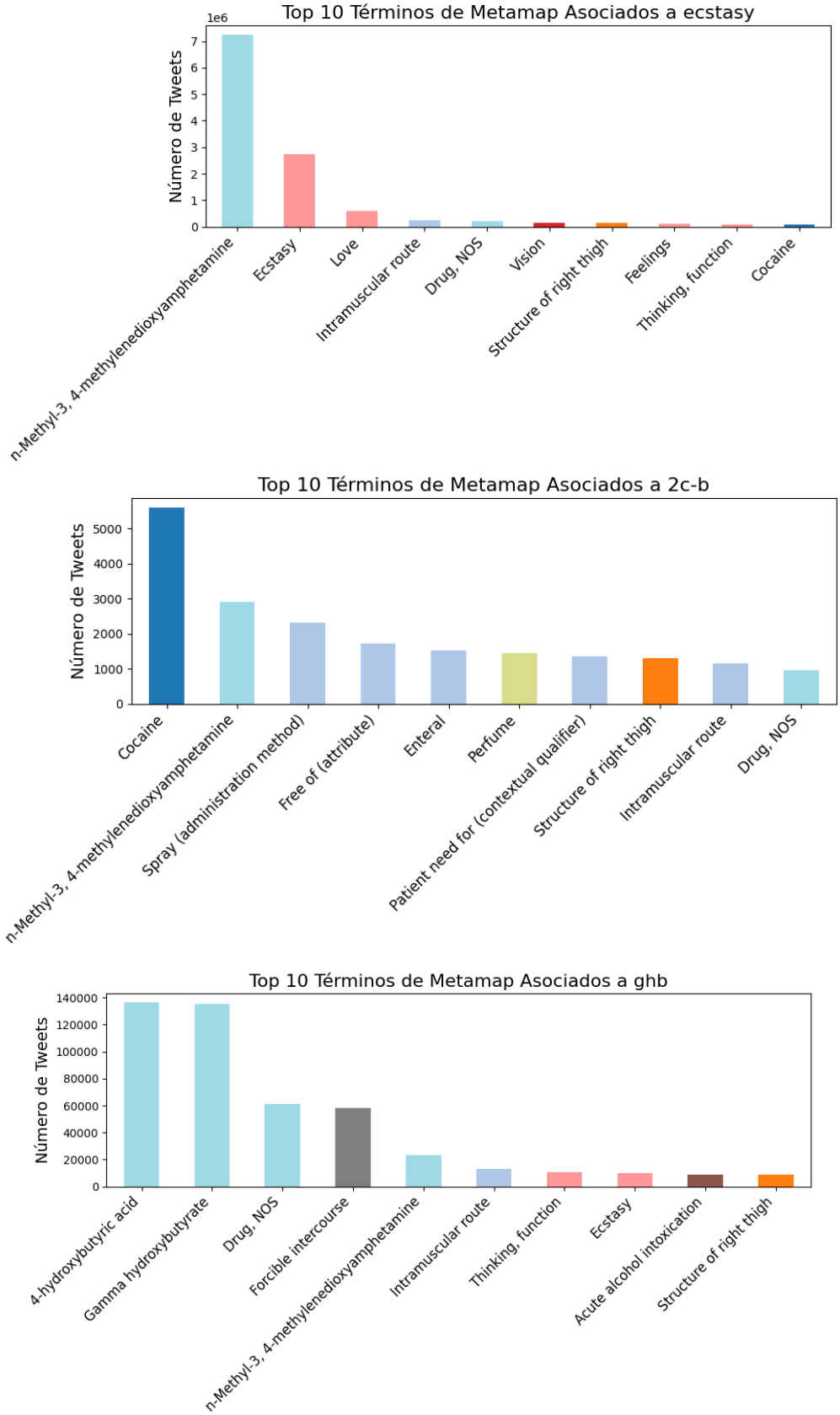


Figura A.16. Distribución de tweets por MetaMap y por droga.

En la **Figura A.17** se muestra los WordCloud de los términos MetaMap por droga.

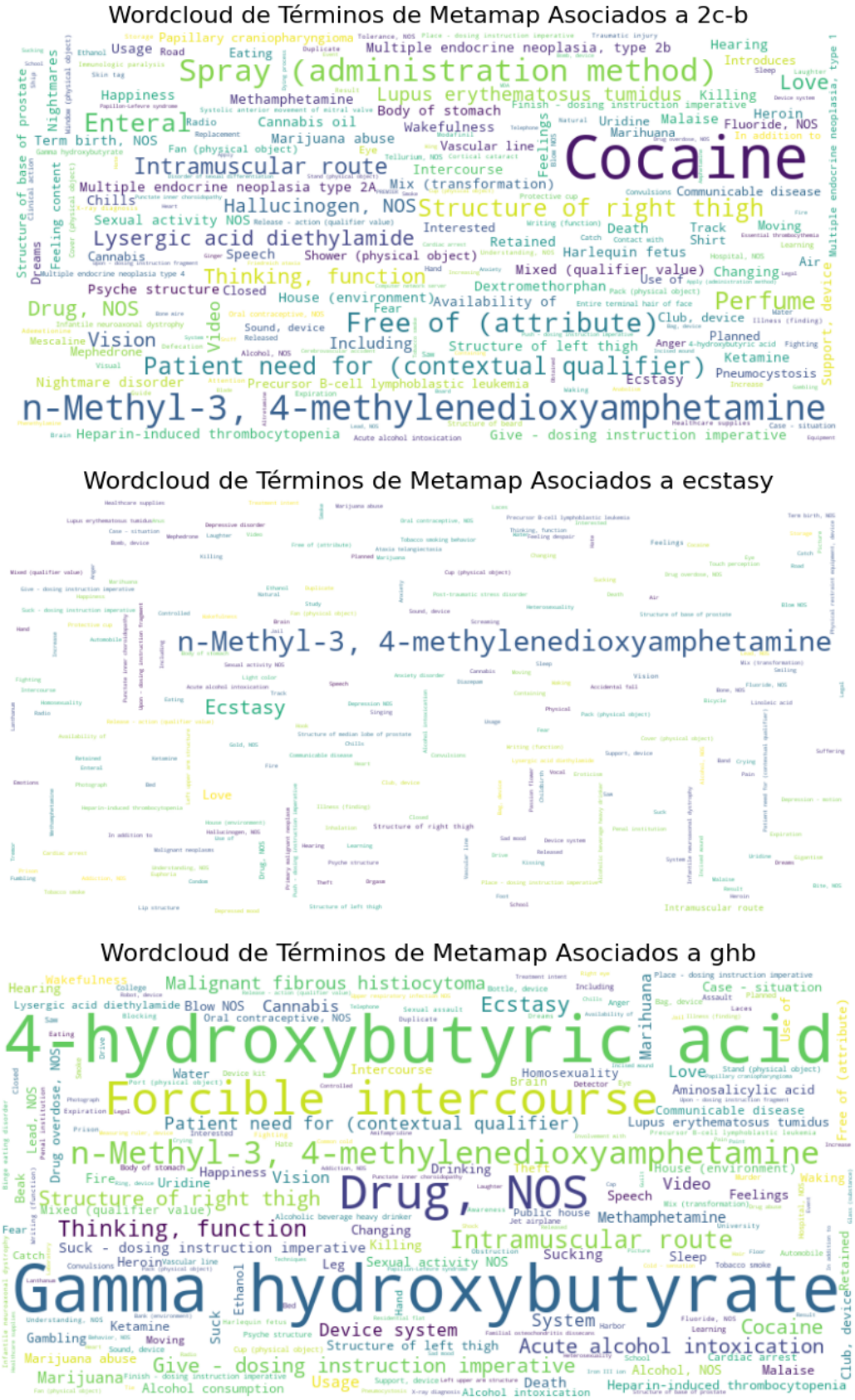


Figura A.17. WordCloud de MetaMap por droga.

Anexo IV: Consulta SQL para crear el *dataset* final

```
SELECT
    t.text,
    t.media,
    CASE WHEN ub.country_code IN ('AL', 'AD', 'AT', 'BY', 'BE', 'BA',
'BG', 'HR', 'CY', 'CZ', 'DK', 'EE', 'FI', 'FR', 'GE', 'DE', 'GR', 'HU',
'IS', 'IE', 'IT', 'KZ', 'LV', 'LI', 'LT', 'LU', 'MT', 'MD', 'MC', 'ME',
'NL', 'MK', 'NO', 'PL', 'PT', 'RO', 'RU', 'SM', 'RS', 'SK', 'SI', 'ES',
'SE', 'CH', 'TR', 'UA', 'GB', 'XK') THEN 1 ELSE 0 END AS is_europe,
    CASE WHEN ub.country_code IN ('DZ', 'AO', 'BJ', 'BW', 'BF', 'BI',
'CM', 'CV', 'CF', 'TD', 'KM', 'CG', 'CD', 'CI', 'DJ', 'EG', 'GQ', 'ER',
'SZ', 'ET', 'GA', 'GM', 'GH', 'GN', 'GW', 'KE', 'LS', 'LR', 'LY', 'MG',
'MW', 'ML', 'MR', 'MU', 'MA', 'MZ', 'NA', 'NE', 'NG', 'RW', 'ST', 'SN',
'SC', 'SL', 'SO', 'ZA', 'SS', 'SD', 'TZ', 'TG', 'TN', 'UG', 'EH', 'ZM',
'ZW') THEN 1 ELSE 0 END AS is_africa,
    CASE WHEN ub.country_code IN ('AS', 'AU', 'CK', 'FJ', 'PF', 'GU',
'KI', 'MH', 'FM', 'NR', 'NC', 'NZ', 'NU', 'MP', 'PW', 'PG', 'PN', 'WS',
'SB', 'TK', 'TO', 'TV', 'VU', 'WF') THEN 1 ELSE 0 END AS is_oceania,
    CASE WHEN ub.country_code IN ('AF', 'AM', 'AZ', 'BH', 'BD', 'BT',
'BN', 'KH', 'CN', 'TL', 'IN', 'ID', 'IR', 'IQ', 'IL', 'JP', 'JO', 'KW',
'KG', 'LA', 'LB', 'MY', 'MV', 'MN', 'MM', 'NP', 'KP', 'OM', 'PK', 'PS',
'PH', 'QA', 'SA', 'SG', 'KR', 'LK', 'SY', 'TW', 'TJ', 'TH', 'TR', 'TM',
'AE', 'UZ', 'VN', 'YE') THEN 1 ELSE 0 END AS is_asia,
    CASE WHEN ub.country_code IN ('AG', 'AR', 'BS', 'BB', 'BZ', 'BO',
'BR', 'CA', 'CL', 'CO', 'CR', 'CU', 'DM', 'DO', 'EC', 'SV', 'GD', 'GT',
'GY', 'HT', 'HN', 'JM', 'MX', 'NI', 'PA', 'PY', 'PE', 'KN', 'LC', 'VC',
'SR', 'TT', 'US', 'UY', 'VE') THEN 1 ELSE 0 END AS is_america,
    CASE WHEN ub.country_code IN ('AE', 'BH', 'CY', 'EG', 'IR', 'IQ',
'IL', 'JO', 'KW', 'LB', 'OM', 'PS', 'QA', 'SA', 'SY', 'TR', 'YE') THEN 1
ELSE 0 END AS is_middle_east,
    CASE WHEN m.tweet_id IS NOT NULL THEN 1 ELSE 0 END as mention,
    CASE WHEN rt.reference_type THEN 1 ELSE 0 END AS reference,
    tm.like_count,
    tm.retweet_count,
    tm.reply_count,
    tm.quote_count,
    u.verified AS user_verified,
    u.followers_count AS user_followers,
    u.following_count AS user_following,
    u.tweet_count AS user_tweet_count,
```

```
    u.listed_count AS user_listed_count,
    CASE WHEN u.location IS NOT NULL THEN 1 ELSE 0 END AS user_location,
    CASE WHEN tw.keyword_id = 1 THEN 1 ELSE 0 END AS mentions_ghb,
    CASE WHEN tw.keyword_id = 14 THEN 1 ELSE 0 END AS mentions_ecstasy,
    CASE WHEN tw.keyword_id = 31 THEN 1 ELSE 0 END AS mentions_2cb,
    ct.classification
FROM classified_tweets ct
LEFT JOIN tweet_metrics tm ON ct.tweet_id = tm.tweet_id
LEFT JOIN tweet_keyword tw ON tw.tweet_id = ct.tweet_id
LEFT JOIN ubication ub ON ub.tweet_id = ct.tweet_id
LEFT JOIN mention m ON m.tweet_id = ct.tweet_id
LEFT JOIN referenced_tweet rt ON rt.tweet_id = ct.tweet_id
LEFT JOIN tweet t ON ct.tweet_id = t.id
LEFT JOIN user u ON t.author_id = u.id
WHERE tw.keyword_id IN (1, 14, 31)
```

Anexo V: Modelos y búsqueda de hiperparámetros

En este anexo se detallan los modelos de clasificación empleados en el entrenamiento, así como los espacios de búsqueda definidos para la optimización de hiperparámetros. La búsqueda se realizó mediante validación cruzada con *RandomizedSearchCV*, con el objetivo de mejorar el rendimiento de cada modelo sin sobreajuste.

A continuación, se muestra el código Python donde se declaran los modelos y los espacios de búsqueda:

```
models = {
    "DT": DecisionTreeClassifier(),
    "RF": RandomForestClassifier(n_estimators=50),
    "AdaBoost": AdaBoostClassifier(),
    "Bagging": BaggingClassifier(),
    "LR": LogisticRegression(max_iter=1000),
    "XGBoost": XGBClassifier(random_state=R, verbosity=0),
    "MLP": MLPClassifier(max_iter=500)
}

hyperparameters = {
    "DT": {'m__splitter': ['best', 'random'],
           'm__max_features': ['sqrt', 'log2'],
           'm__criterion': ['gini', 'entropy', 'log_loss']},

    "RF": {'m__n_estimators': randint(100, 250),
           'm__max_features': ['sqrt', 'log2'],
           'm__criterion': ['gini', 'entropy']},

    "Bagging": {'m__n_estimators': randint(10, 100),
                'm__max_samples': [0.8, 1.0],
                'm__max_features': [0.8, 1.0],
                'm__warm_start': [True, False]},

    "AdaBoost": {'m__n_estimators': randint(50, 150),
                 'm__learning_rate': uniform(0.8, 1.2)},

    "LR": {'m__penalty': ['l1', 'l2', 'elasticnet', None],
           'm__solver': ['lbfgs', 'sag', 'saga', 'newton-cholesky']},

    "XGBoost": {'m__max_depth': [2, 3, 5, 7, 10],
                'm__n_estimators': [10, 100, 500]},

    "MLP": {'m__activation': ['identity', 'logistic', 'tanh', 'relu'],
            'm__hidden_layer_sizes': randint(50, 150),
            'm__learning_rate': ['constant', 'invscaling', 'adaptive']}
}
```

Anexo VI: Resultados de validación cruzada para las estrategias de entrenamiento evaluadas

Tabla A.3. Tabla de métricas de validación cruzada para las cuatro estrategias de entrenamiento

	Método	F1	REC	PREC	ACC
COST-SENSITIVE	DT	0,733±0,020	0,976±0,002	0,753±0,015	0,745±0,024
	RF	0,842±0,009	0,986±0,001	**0,987±0,008**	0,739±0,016
	LR	0,581±0,004	0,931±0,001	0,420±0,003	**0,942±0,012**
	XGBoost	**0,894±0,013**	**0,990±0,001**	0,931±0,012	0,860±0,020
	AdaBoost	0,795±0,018	0,981±0,002	0,867±0,027	0,735±0,024
	Bagging	0,852±0,014	0,986±0,001	0,968±0,012	0,758±0,021
	MLP	0,883±0,010	0,989±0,001	0,930±0,015	0,852±0,017
SMOTE_PRE_CV	DT	0,970±0,001	0,970±0,001	0,964±0,001	0,976±0,001
	RF	0,993±0,000	0,993±0,001	0,991±0,001	0,996±0,000
	LR	0,953±0,002	0,953±0,002	0,949±0,006	0,957±0,007
	XGBoost	0,990±0,001	0,990±0,001	0,986±0,000	0,994±0,000
	AdaBoost	0,949±0,001	0,948±0,001	0,943±0,002	0,954±0,003
	Bagging	**0,996±0,000**	**0,996±0,000**	**0,994±0,001**	0,998±0,000
	MLP	0,995±0,000	0,995±0,000	0,990±0,001	**0,999±0,000**
SMOTE_IN_CV	DT	0,672±0,010	0,957±0,001	0,569±0,015	0,821±0,022
	RF	0,856±0,009	0,985±0,001	0,848±0,012	0,866±0,022
	LR	0,644±0,022	0,949±0,005	0,502±0,026	0,898±0,014
	XGBoost	0,828±0,007	0,982±0,001	0,795±0,013	0,859±0,015
	AdaBoost	0,622±0,011	0,943±0,002	0,469±0,008	**0,925±0,018**
	Bagging	**0,886±0,015**	**0,988±0,002**	**0,897±0,018**	0,876±0,020
	MLP	0,861±0,010	0,986±0,001	0,852±0,021	0,871±0,013
SIN_OVERSAMPLING	DT	0,727±0,009	0,973±0,002	0,726±0,020	0,744±0,010
	RF	0,845±0,015	0,986±0,001	0,736±0,019	0,986±0,008
	LR	0,815±0,006	0,983±0,001	0,745±0,009	0,898±0,004
	XGBoost	**0,887±0,012**	**0,989±0,001**	0,829±0,019	**0,953±0,003**
	AdaBoost	0,786±0,025	0,980±0,002	0,725±0,024	0,857±0,025
	Bagging	0,849±0,013	0,986±0,001	0,754±0,019	0,971±0,006
	MLP	0,874±0,006	0,988±0,001	**0,836±0,013**	0,923±0,010

Este documento esta firmado por

	Firmante	CN=tfgm.fi.upm.es, OU=CCFI, O=ETS Ingenieros Informaticos - UPM, C=ES
	Fecha/Hora	Tue Jun 03 20:42:05 CEST 2025
	Emisor del Certificado	EMAILADDRESS=camanager@etsiinf.upm.es, CN=CA ETS Ingenieros Informaticos, O=ETS Ingenieros Informaticos - UPM, C=ES
	Numero de Serie	561
	Metodo	urn:adobe.com:Adobe.PPKLite:adbe.pkcs7.sha1 (Adobe Signature)