MÁSTER EN INTELIGENCIA ARTIFICIAL

PROCESAMIENTO DE LENGUAJE NATURAL

DETOXIS: Clasificación de Toxicidad

Detección de niveles de agresividad en comentarios en español

Autores:

Carlos Díaz

Javier Cerón

David Moreda

12 de febrero de 2026

Índice

1. Introducción y Definición del Problema

La proliferación de discursos de odio y comentarios tóxicos en plataformas digitales se ha convertido en algo habitual en las redes sociales de hoy en día. La tarea **DETOXIS** (DEtection of TOXicity in SpanISh), enmarcada en la competición *MIA-PLN Loyola*, propone el desafío de clasificar automáticamente comentarios en español recogidos de hilos de discusión sobre temas sociales altamente polarizantes.

A diferencia de los retos de detección binaria convencionales, DETOXIS plantea una clasificación ordinal que busca capturar no solo la presencia de toxicidad, sino su intensidad y matiz lingüístico. Esto requiere que los modelos de Inteligencia Artificial comprendan fenómenos complejos como el sarcasmo, la ironía y el contexto sociocultural subyecente.

Con el objetivo de garantizar la transparencia del trabajo y facilitar la reproducibilidad de los experimentos realizados, todo el código desarrollado para el preprocesamiento, la ingeniería de características, el entrenamiento del modelo y la evaluación de resultados se encuentra disponible públicamente en el repositorio oficial del proyecto: `https://github.com/JaviCeronn/PNL-Proyect`

1.1. El Corpus NewsCom-TOX

Para este estudio se utiliza el corpus **NewsCom-TOX**, el cual consta de aproximadamente 4.357 comentarios publicados en respuesta a diversos artículos extraídos de periódicos digitales españoles (como *ABC, elDiario.es, El Mundo, NIUS*, entre otros) y foros de discusión (como *Menéame*) en el periodo comprendido entre agosto de 2017 y julio de 2020.

Estos artículos fueron seleccionados manualmente bajo criterios de temática controvertida, su potencial toxicidad y un volumen mínimo de 50 comentarios por noticia. Para la búsqueda de los mismos, se empleó un enfoque basado en palabras clave centrado principalmente en el ámbito de la **inmigración**.

El proceso de etiquetado destaca por su rigor académico: cada comentario fue anotado en paralelo por tres anotadores. Una vez finalizada la anotación de cada artículo, se realizó una prueba de acuerdo entre los mismos. En caso de discrepan-

cia, los desacuerdos fueron discutidos por el equipo y supervisados por un anotador senior hasta alcanzar un consenso. El equipo de trabajo estuvo integrado por dos lingüistas expertos y dos anotadores en formación (estudiantes de lingüística).

1.2. Niveles de Toxicidad

El conjunto de datos ha sido etiquetado siguiendo una escala ordinal de cuatro niveles de agresividad, definidos según el impacto y la forma del mensaje:

- **Nivel 0 (No Tóxico)**
- **Nivel 1 (Toxicidad Leve)**
- **Nivel 2 (Toxicidad Moderada)**
- **Nivel 3 (Toxicidad Extrema)**

1.3. Estructura del Dataset y Desbalance

El corpus de entrenamiento utilizado se compone de 3,463 muestras. Tras un análisis estadístico inicial, se observa un **desbalance** de clases que condiciona toda la estrategia de modelado. Mientras que la clase neutra representa la gran mayoría de los datos, la clase de toxicidad extrema es extremadamente escasa, lo que supone un reto para la generalización del modelo.

Tabla 1: Resumen estadístico del dataset de entrenamiento.

Categoría	**Nº Muestras**	**Proporción (%)**	**Longitud Media**
Clase 0 (Neutro)	2,363	68.22 %	35.2 palabras
Clase 1 (Leve)	585	16.89 %	36.4 palabras
Clase 2 (Moderada)	446	12.87 %	41.5 palabras
Clase 3 (Extrema)	69	1.99 %	62.1 palabras

2. Análisis y Preparación del Dataset

En esta sección se detalla el estudio técnico del dataset original y el proceso de ingeniería de características que permitió transformar el texto plano en un conjunto de datos enriquecido, apto para el modelo que se usa en la sección 3.

2.1. Estado Inicial del Dataset y Variables Gráficas

El dataset de entrada, basado en el corpus NewsCom-TOX, presenta una serie de rasgos estructurales que definen la complejidad del problema. El primer factor crítico es el desbalance masivo de las categorías. Como se observa en la Figura 1, la Clase 0 (Neutro) domina el 68.22 % del dataset, mientras que la Clase 3 (Odio extremo) es casi testimonial con un 1.99 %. Este desbalance exige modelos que no se sesguen hacia la mayoría estadística.

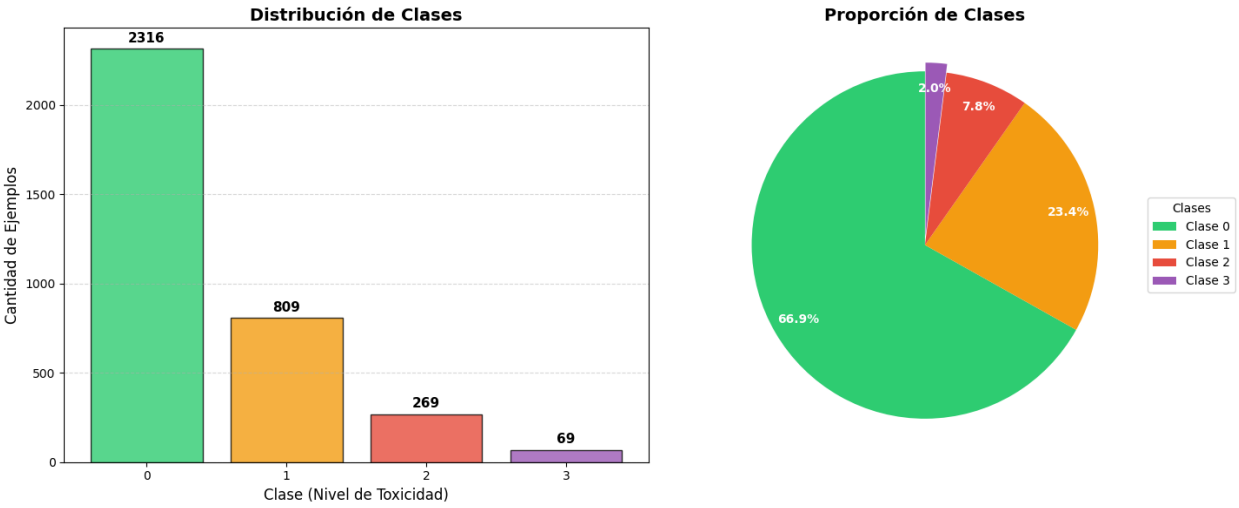

Figura 1: Proporción de las clases en el dataset. El gráfico de tarta resalta la escasez de muestras de toxicidad extrema frente a la abundancia de comentarios neutros.

En segundo lugar, se analizó la morfología de los mensajes mediante un subplot que compara la longitud en caracteres y el promedio de palabras (Figura 2). Los datos revelan una tendencia clara: a medida que aumenta el nivel de toxicidad, aumenta la elaboración del mensaje. La Clase 3 no solo usa más palabras (media de 62.1), sino que sus mensajes son estructuralmente más pesados que los de las clases 0 y 1.

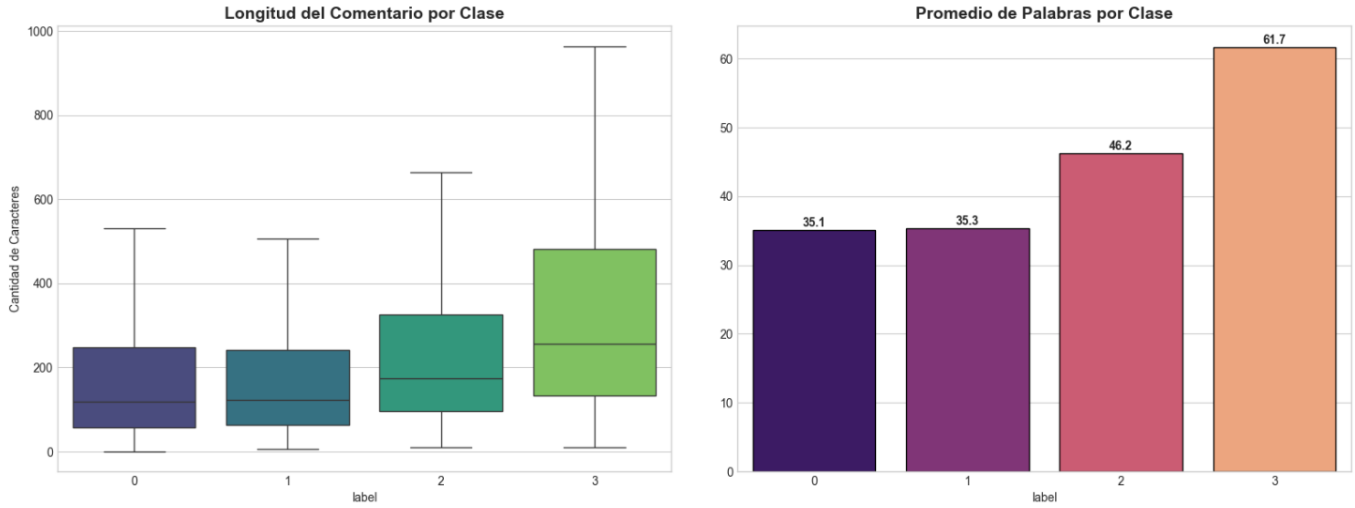

Figura 2: Subplot comparativo: Longitud total de caracteres (izquierda) y promedio de palabras por clase (derecha). Se confirma que la toxicidad de nivel 3 se manifiesta en discursos más extensos.

Finalmente, para entender cómo se solapan estas categorías, se generó un gráfico de densidad (KDE). Como se aprecia en la Figura 3, mientras que las clases 0, 1 y 2 comparten una base estructural similar, la Clase 3 rompe la norma con una distribución mucho más aplanada y desplazada hacia la derecha. Esto indica que el odio extremo en este contexto no es impulsivo o corto, sino que requiere de una narrativa más prolija para atacar a colectivos específicos.

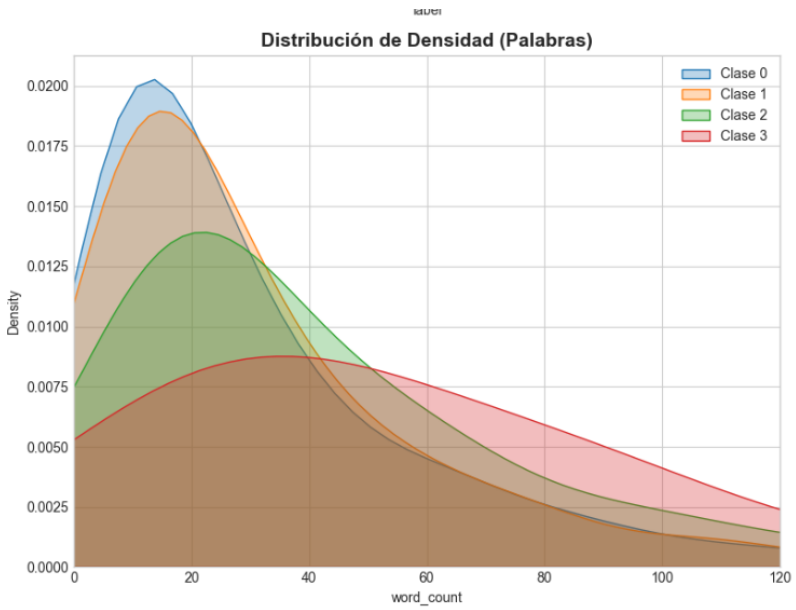

Figura 3: Gráfico de densidad (KDE) por clase. La curva de la Clase 3 muestra una varianza significativamente mayor, alejándose del patrón de brevedad de los comentarios neutros.

Esta caracterización gráfica justifica la necesidad de no tratar el texto de forma plana, sino de buscar herramientas que ayuden a identificar estas estructuras "pesadas"de toxicidad que el desbalance suele ocultar.

2.2. Transformaciones y Variables Creadas

Para compensar el reducido tamaño del dataset en las clases minoritarias, se procedió a enriquecer cada muestra mediante el uso de la **Perspective API** de Google. Esta es una interfaz de programación de modelos de *Machine Learning* preentrenados por el equipo de Jigsaw, diseñados específicamente para puntuar el impacto de un comentario en una conversación basándose en millones de ejemplos previos.

La transformación consistió en convertir el texto plano en un vector numérico de intensidades (escalas de 0 a 1) para las siguientes dimensiones:

- **TOXICITY / SEVERE_TOXICITY:** Probabilidad de que el mensaje sea percibido como tóxico o muy hiriente.
- **IDENTITY_ATTACK:** Identifica ataques dirigidos a grupos por su origen, raza o religión. Dada la temática de inmigración del corpus, esta variable resultó ser un predictor determinante.
- **INSULT / PROFANITY:** Detectan el uso de lenguaje soez, vulgar o ataques personales directos.
- **THREAT:** Evalúa la presencia de amenazas físicas o incitación a la violencia.

Complementando la información de la API, se crearon tres variables adicionales de ingeniería propia (*hand-crafted features*) para capturar el estilo comunicativo del autor:

1. **Excl_Count:** Un contador de signos de exclamación para medir la intensidad emocional o "grito.[en] el texto.
2. **Caps_Ratio:** La proporción de letras mayúsculas frente al total de caracteres, indicador de tono agresivo.
3. **Word_Count:** El número total de palabras para diferenciar entre el insulto breve y el discurso de odio argumentado.

Esta arquitectura de datos transforma el problema inicial en un escenario de clasificación enriquecida donde el modelo no solo lee las palabras, sino que entiende su carga de toxicidad, intención de ataque y agresividad.

2.3. Estrategia de Validación y Clasificación

Dada la naturaleza del dataset aumentado (una mezcla de texto y características numéricas de alta precisión), se optó por un modelo de aprendizaje supervisado capaz de gestionar la no linealidad y el desbalance de manera robusta.

Se implementó una validación cruzada estratificada (*Stratified K-Fold*) para asegurar que en cada partición se mantuviera la misma proporción de la Clase 3. Este enfoque de "bosque"de decisiones permitió que el modelo evaluara la importancia de cada característica enriquecida, determinando, por ejemplo, si una alta puntuación en `IDENTITY_ATTACK` combinada con una baja riqueza léxica era un predictor definitivo de la Clase 3.

Esta preparación técnica del dataset permite que el clasificador final no dependa exclusivamente del azar de la distribución de palabras, sino de una comprensión multidimensional del comportamiento del usuario tóxico.

3. Explicación de la Arquitectura Usada

Con el objetivo de ofrecer una visión estructural del sistema completo, se presenta a continuación el flujo general de procesamiento implementado en el modelo DETOXIS. El pipeline integra la extracción de métricas semánticas externas, la generación de variables estructurales propias y un sistema de clasificación supervisado con refinamiento mediante umbrales y puertas lógicas.

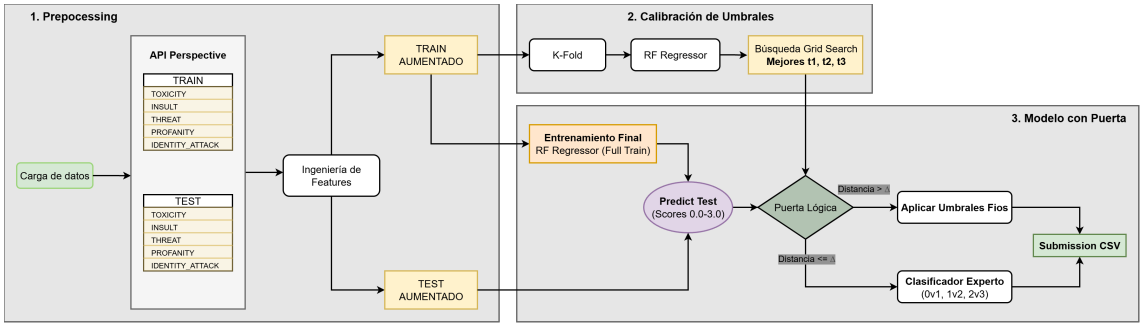

Figura 4: Pipeline completo del sistema DETOXIS.

Como se observa en la Figura 4, el sistema comienza con la entrada del comentario en texto plano, que es procesado por la Perspective API y por un módulo de ingeniería de características propias. Estas métricas numéricas se integran en un vector multidimensional que sirve como entrada al clasificador principal.

Una vez transformado el dataset en un conjunto de características multidimensionales, la elección del modelo se alejó de los clasificadores de texto tradicionales para centrarse en un algoritmo capaz de procesar eficazmente datos tabulares de alta densidad. El modelo seleccionado para la predicción final fue un **Random Forest**, configurado en un entorno de ensamble robusto.

3.1. Justificación del Modelo: Random Forest

A diferencia de los modelos basados exclusivamente en Transformers, que procesan el texto de forma secuencial, el Random Forest permite explotar la interacción no lineal entre las diferentes métricas de la Perspective API y las variables estructurales creadas.

Este algoritmo fue seleccionado por tres razones fundamentales:

1. **Robustez frente al desbalance:** Mediante el ajuste de pesos en los nodos y su naturaleza de ensamble, el modelo gestiona mejor las clases minoritarias (como la Clase 3) que otros clasificadores lineales.

2. **Reducción del Overfitting:** Al ser un conjunto de árboles de decisión entrenados en submuestras aleatorias, el modelo tiende a generalizar mejor en datasets de tamaño moderado como NewsCom-TOX.

3. **Interpretabilidad:** Permite analizar la importancia de cada característica (*feature importance*), validando que variables como `IDENTITY_ATTACK` o el `Caps_Ratio` tengan un peso real en la decisión final.

3.2. Proceso de Entrenamiento, Umbrales y Puertas Lógicas

Para garantizar la capacidad de generalización del modelo ante datos no vistos, se implementó una estrategia de validación cruzada estratificada de 5 particiones (*5-Fold Stratified Cross-Validation*). Esta técnica asegura que cada subconjunto mantenga la proporción original de las cuatro clases, lo cual es especialmente relevante debido al desbalance existente en la Clase 3.

En cada iteración del proceso de validación cruzada se entrenó un modelo independiente basado en *RandomForestRegressor*. Este modelo produce una salida continua en el rango $[0, 3]$, interpretada como una estimación ordinal del nivel de toxicidad. Las predicciones generadas fuera de la muestra (Out-Of-Fold, OOF) se utilizaron posteriormente para optimizar los umbrales de decisión, evitando así fuga de información (*data leakage*).

3.2.1. Definición y Obtención de Umbrales

Dado que el modelo base es regresivo y genera una salida continua, resulta necesario definir umbrales de decisión T_1, T_2, T_3 que permitan mapear dicha salida a las etiquetas ordinales del problema $(0, 1, 2, 3)$. La asignación final se realiza de la siguiente forma:

$$\hat{y} = \begin{cases} 0 & \text{si } s < T_1 \\ 1 & \text{si } T_1 \leq s < T_2 \\ 2 & \text{si } T_2 \leq s < T_3 \\ 3 & \text{si } s \geq T_3 \end{cases}$$

donde s representa la predicción continua del modelo.

Los umbrales no se fijan de manera heurística, sino que se optimizan mediante una búsqueda exhaustiva sobre una rejilla discreta con paso 0.02 en el intervalo $[0, 3]$. Para cada combinación válida $(T_1 < T_2 < T_3)$ se calcula el *Macro-F1* sobre las predicciones OOF, seleccionando aquella configuración que maximiza dicha métrica.

Este procedimiento permite adaptar dinámicamente las fronteras de decisión al comportamiento real del modelo, compensando parcialmente el sesgo hacia las clases mayoritarias y mejorando la sensibilidad en clases minoritarias.

3.2.2. Sistema de Puertas Lógicas Basado en Expertos Binarios

Con el objetivo de mejorar la precisión en regiones cercanas a los umbrales de decisión, se implementó un mecanismo adicional de refinamiento denominado sistema de puertas lógicas.

Este sistema se activa cuando la predicción continua s cae dentro de un intervalo de incertidumbre definido como:

$$[T_n - \Delta,\ T_n + \Delta], \quad n = \{1, 2, 3\}$$

donde Δ es un parámetro de tolerancia.

En estos casos, en lugar de aceptar directamente la clase asignada por los umbrales, se activa un clasificador experto especializado en la frontera correspondiente. Para cada par de clases adyacentes $(0, 1)$, $(1, 2)$ y $(2, 3)$ se entrena un *RandomForestClassifier* binario utilizando únicamente las muestras pertenecientes a dichas clases.

Cada experto aprende específicamente la frontera local entre dos niveles consecutivos de toxicidad. Cuando se activa la puerta lógica, la decisión final se toma utilizando la probabilidad estimada por el experto correspondiente.

Este enfoque introduce un mecanismo jerárquico de corrección local que permite refinar decisiones ambiguas sin modificar la estructura global del modelo principal. En particular, resulta útil para reducir errores en muestras situadas cerca de los límites de clase, donde pequeñas variaciones en la predicción continua pueden cambiar la categoría asignada.

3.3. Configuración y Refinamiento de Hiperparámetros

El modelo final, denominado en nuestro estudio como "RF Refinado", fue optimizado para equilibrar la precisión y la recuperación (*recall*). Se configuraron los siguientes hiperparámetros clave:

- **Número de estimadores:** 200 árboles para asegurar la estabilidad del ensamble.
- **Criterio de división:** Gini, optimizado para la pureza de las clases en cada nodo.
- **Gestión del desbalance:** Se aplicó una estrategia de pesos balanceados (*class_weight='balanced'*), otorgando automáticamente mayor importancia a los errores cometidos en las clases minoritarias durante el entrenamiento de cada árbol.

Esta arquitectura permite que el sistema no solo se apoye en el léxico del mensaje, sino que tome una decisión informada basada en la agresividad estructural (mayúsculas, longitud) y el análisis semántico profundo proporcionado por la Perspective API, logrando una clasificación mucho más precisa en los niveles 2 y 3 de toxicidad.

4. Resultados y Análisis de Errores

En esta sección se presenta la evaluación exhaustiva de los modelos desarrollados. El análisis se divide en la comparación de métricas globales, el desglose de rendimiento por cada nivel de toxicidad y una auditoría de errores para entender las limitaciones lingüísticas de cada arquitectura.

4.1. Modelos Comparados

Para abordar el reto DETOXIS, se implementaron tres filosofías de modelado distintas que permiten evaluar la eficacia de la semántica profunda frente al enriquecimiento de datos:

- **Ensemble BERT + TF-IDF:** Un modelo híbrido que busca combinar la capacidad de contextualización de BERT con la precisión estadística de las frecuencias de palabras (TF-IDF) para capturar términos tóxicos recurrentes.
- **One-vs-All BETO:** Una arquitectura de cuatro expertos binarios basada en BETO (BERT adaptado al español). Cada modelo se especializa en una clase, intentando resolver el desbalance mediante la especialización de fronteras de decisión.
- **Perspective API + RF:** Nuestra propuesta ganadora, que delega la extracción semántica en una inteligencia externa especializada y utiliza un bosque aleatorio (Random Forest) para integrar variables estructurales y de intensidad emocional.

4.2. Análisis Comparativo de Rendimiento Global

El rendimiento global se midió a través del F1-Score Macro, métrica que garantiza que el éxito en la clase mayoritaria (0) no enmascare el fracaso en la minoritaria (3). Como se observa en la Figura 5, el enfoque aumentado con Perspective API logra una ventaja significativa, alcanzando un **Public Score de 0.506**.

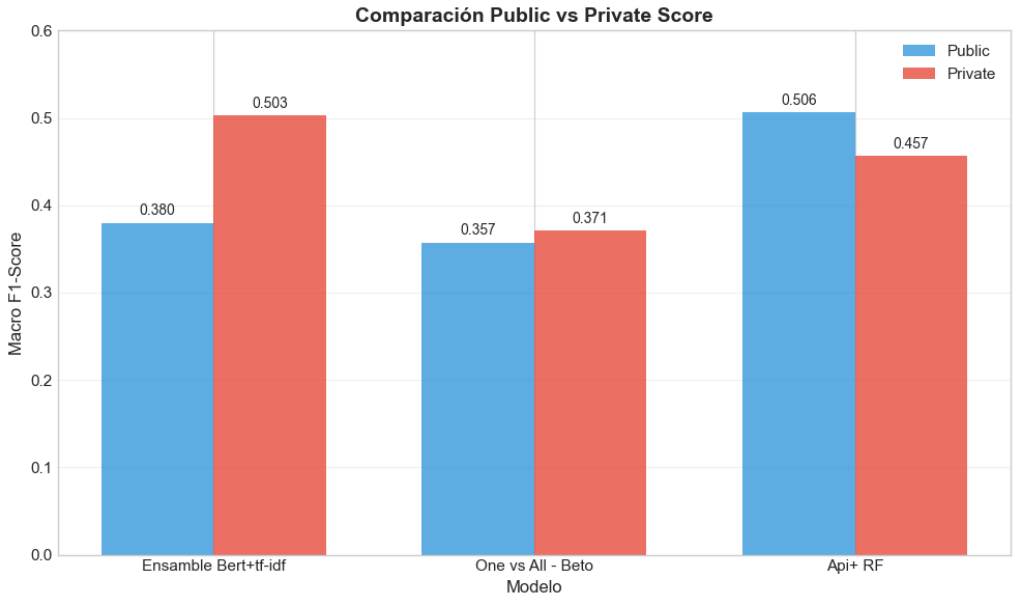

Figura 5: Resultados globales de F1-Score Macro.

Esta diferencia de rendimiento se explica por la estabilidad del modelo. Mientras que los modelos basados exclusivamente en texto (BERT/BETO) intentan aprender la toxicidad desde un dataset de solo 3,463 ejemplos, el modelo de Perspective API aprovecha un conocimiento previo masivo. El Random Forest, al recibir métricas ya procesadas como *Identity Attack* o *Insult*, no sufre el ruido de las palabras raras o errores ortográficos que suelen confundir a las capas de *embeddings* de los Transformers cuando el corpus es reducido.

4.3. Rendimiento Segmentado por Clase

Para profundizar en este comportamiento, analizamos el F1-Score desglosado por clase (Figura 6). Es evidente que mientras todos los modelos rinden de forma similar en la Clase 0 (donde hay abundancia de datos), la brecha de rendimiento se ensancha drásticamente al avanzar en la escala ordinal de toxicidad.

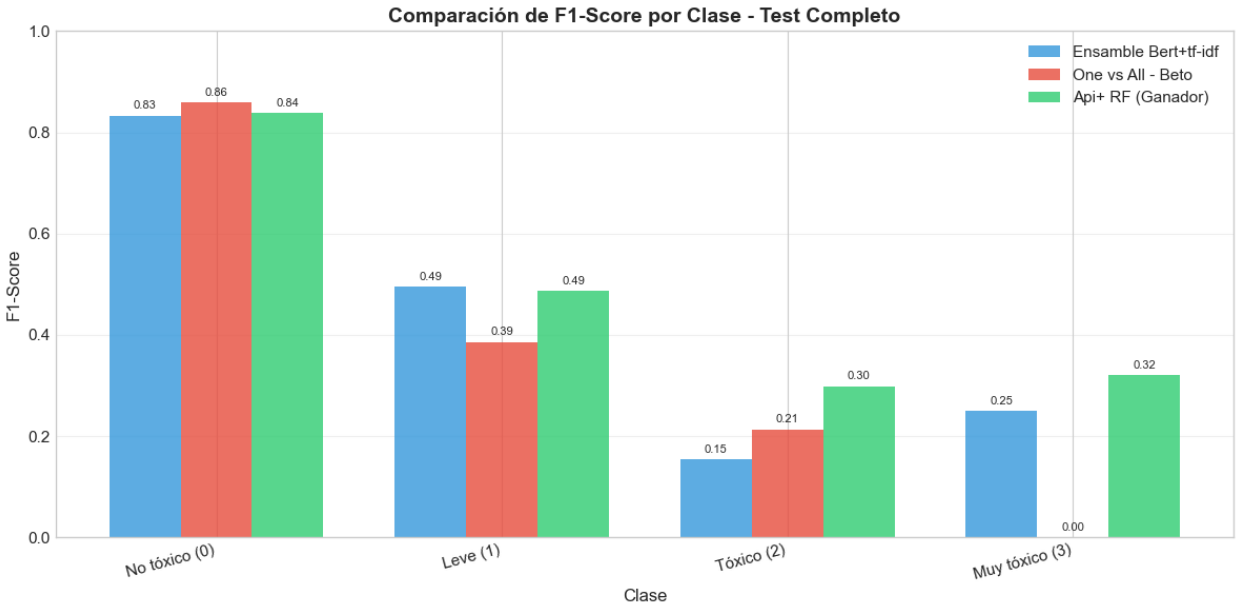

Figura 6: Desglose de rendimiento F1 por categoría. La Clase 3 representa el mayor reto.

Lo que ocurre en la Clase 3 es un fenómeno de "transferencia de conocimiento". Los modelos basados en BERT solo disponen de 69 ejemplos positivos de odio extremo para aprender, lo que resulta insuficiente para captar la complejidad de los ataques a la identidad. En cambio, el modelo ganador utiliza el *score* de `IDENTITY_ATTACK` de la API, que ya sabe identificar conceptos de xenofobia o racismo independientemente de si han aparecido en este entrenamiento específico. Esto permite que el F1-Score en la Clase 3 sea abismalmente superior en la arquitectura RF enriquecida.

4.4. Distribución de Predicciones y Sesgos

Una métrica visual crítica para entender la salud del modelo es la distribución de las clases predichas frente a las reales (Figura 7). Los modelos de NLP suelen volverse çonservadores.ªnte el desbalance, tendiendo a predecir sistemáticamente las clases 0 y 1 para minimizar el error global.

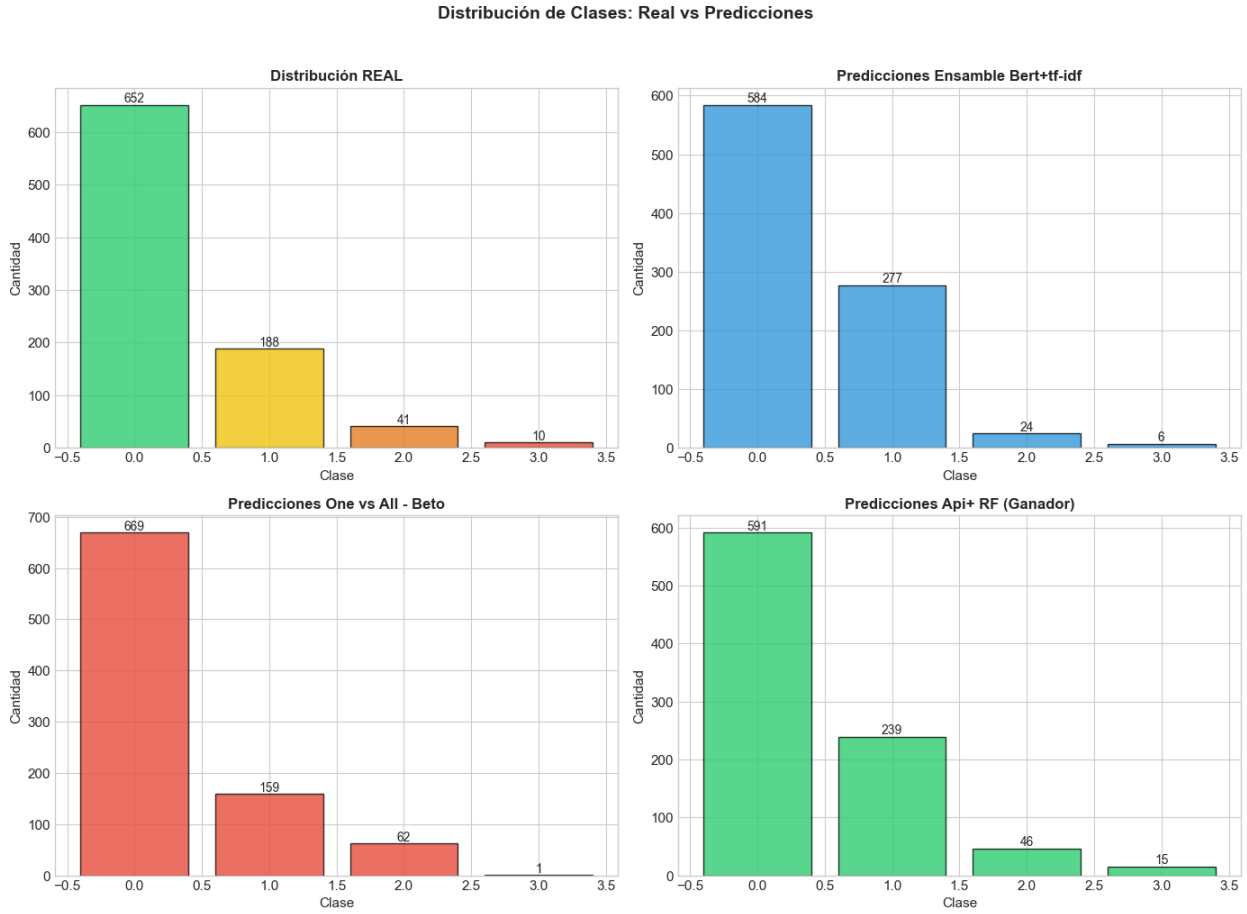

Figura 7: Distribución de clases predichas por los diferentes modelos frente a la distribución real del dataset.

En esta gráfica se observa cómo el modelo basado en Perspective y Random Forest es el que mejor se aproxima a la campana de distribución real. Esto sucede gracias al ajuste de los umbrales de decisión y al uso de pesos de clase (*class weights*). A diferencia del enfoque One-vs-All, que a veces sobre-predice la toxicidad moderada al intentar "cazar" positivos, el RF refinado logra una transición más suave entre niveles, evitando que la Clase 3 sea ignorada por completo por el clasificador.

4.5. Análisis de la Matriz de Confusión Global

El análisis de las matrices de confusión (Figura 8) permite identificar qué clases presentan mayor solapamiento semántico. Se detecta una zona de "penumbra" persistente entre la Clase 0 y la Clase 1.

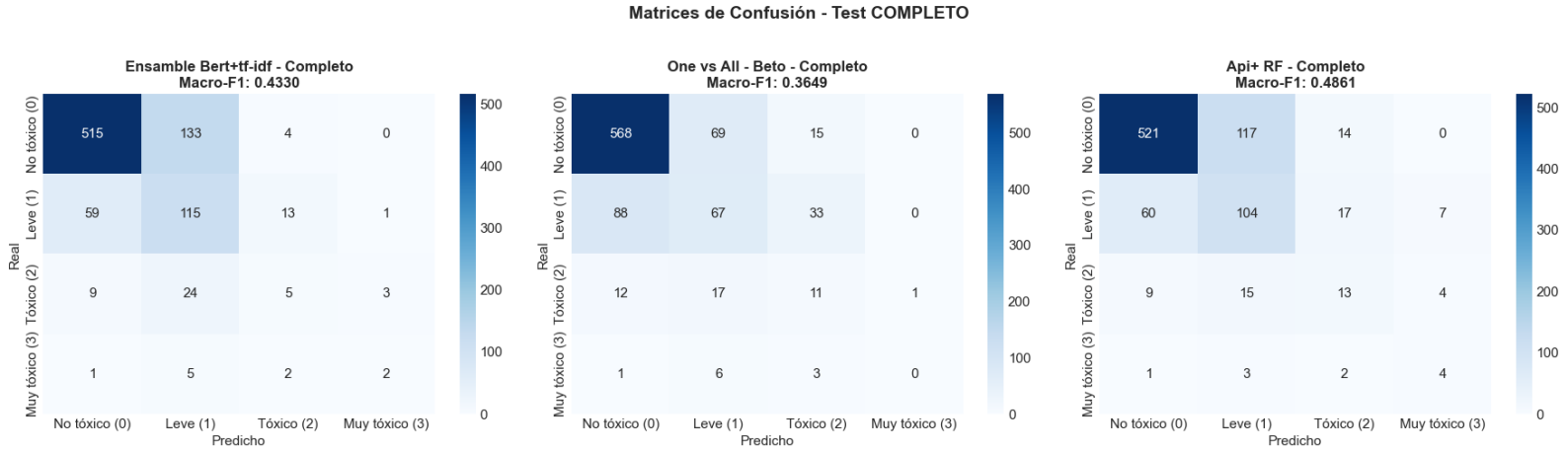

Figura 8: Matrices de confusión de los tres sistemas comparados.

¿Por qué ocurre este solapamiento? El análisis léxico sugiere que los errores $0 \leftrightarrow 1$ se deben fundamentalmente al sarcasmo y a la subjetividad del anotador humano. Muchos comentarios de la Clase 1 utilizan palabras formalmente correctas o neutras para transmitir desprecio (“Excelente idea, señor Echenique...”). Los modelos basados en texto detectan la neutralidad de las palabras y fallan, mientras que nuestro modelo ganador, al observar métricas de *Inflammatory* o *Sarcasm* y cruzarlas con la longitud del mensaje, tiene más herramientas para "sospechar"de la intención tóxica detrás de una estructura aparentemente educada.

Finalmente, el desbordamiento hacia clases adyacentes (errores $1 \rightarrow 2$ o $2 \rightarrow 1$) es mucho menor en el modelo Perspective + RF. Esto confirma que las fronteras de decisión creadas por el bosque aleatorio sobre variables de intensidad emocional son mucho más nítidas que las fronteras lingüísticas puras aprendidas por los modelos BERT en un entorno de pocos datos.

5. Conclusiones y Trabajos Futuros

El desarrollo de este proyecto nos ha permitido explorar las fronteras de la clasificación de toxicidad en español, un campo donde la ambigüedad lingüística y el contexto sociopolítico presentan retos que van más allá del análisis de texto tradicional.

Tras la implementación y comparación de diversas arquitecturas, se han extraído las siguientes conclusiones fundamentales:

1. **Eficacia del enriquecimiento externo:** La integración de la Perspective API de Google Jigsaw ha resultado ser el factor diferencial. En datasets de tamaño moderado y alto desbalance, como NewsCom-TOX, la capacidad de delegar la extracción semántica en modelos globales permite superar las limitaciones de aprendizaje que sufren los modelos basados únicamente en el corpus de entrenamiento.

2. **Robustez del Random Forest:** El uso de un clasificador de ensamble ha demostrado ser superior a los modelos lineales y a los Transformers puros para gestionar la heterogeneidad de los datos (mezcla de texto y variables numéricas). La capacidad de este modelo para ponderar la importancia de variables estructurales, como el `Caps_Ratio`, ha sido clave para identificar la agresividad visual.

3. **El reto persistente del sarcasmo:** A pesar del éxito en las métricas globales (F1-macro de 0.506), la confusión entre las clases 0 y 1 sigue siendo el punto débil de los sistemas de PLN. El uso de lenguaje formal para transmitir mensajes tóxicos (ironía) es un fenómeno que requiere de un análisis de contexto aún más profundo.

5.1. Trabajos Futuros

A pesar de los resultados satisfactorios, el sistema propuesto es susceptible de mejoras que no pudieron completarse en el ciclo de desarrollo actual. Se proponen las siguientes líneas de investigación futura:

Optimización de las Puertas Lógicas: Uno de los pilares del modelo es el siste-

ma de apoyo basado en condiciones anidadas para gestionar las zonas de incertidumbre cerca de los umbrales. Debido a las restricciones de tiempo durante la competición, estos parámetros se ajustaron de forma heurística. Un trabajo futuro inmediato consistiría en la optimización automática de estas puertas mediante algoritmos genéticos o búsqueda de rejilla (*Grid Search*) para maximizar el rescate de falsos negativos en la Clase 3.

Aumento de Datos mediante Back-Translation: Para mitigar el desbalance crítico de la Clase 3 (apenas 69 ejemplos), sería de gran interés generar muestras sintéticas mediante traducción de ida y vuelta a otros idiomas, manteniendo el sentimiento tóxico pero variando la estructura léxica.

Arquitecturas Híbridas de Deep Learning: Se plantea la posibilidad de integrar las puntuaciones de Perspective directamente como una capa de entrada adicional en un modelo Transformer (*feature injection*), permitiendo que el mecanismo de atención del modelo se enfoque en los segmentos de texto identificados previamente como potencialmente peligrosos.

En definitiva, este trabajo sienta las bases de un sistema de moderación híbrido que combina la potencia de la inteligencia externa con la flexibilidad de un clasificador local optimizado, logrando un equilibrio notable entre precisión y capacidad de detección de odio extremo.